
For Reference
Not to be taken
from this library

WITHDRAWN

WITHDRAWN

The New Encyclopædia Britannica

Volume 17

MACROPÆDIA

Knowledge in Depth

FOUNDED 1768
15TH EDITION

Encyclopædia Britannica, Inc.
Jacob E. Safra, Chairman of the Board
Jorge Aguilar-Cauz, President

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

SAN BRUNO PUBLIC LIBRARY

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

Eleventh Edition
© 1911
By Encyclopædia Britannica, Inc.

Twelfth Edition
© 1922
By Encyclopædia Britannica, Inc.

Thirteenth Edition
© 1926
By Encyclopædia Britannica, Inc.

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

Fifteenth Edition
© 1974, 1975, 1976, 1977, 1978, 1979, 1980, 1981, 1982, 1983, 1984, 1985, 1986, 1987, 1988, 1989, 1990, 1991, 1992, 1993, 1994, 1995, 1997, 1998, 2002, 2003, 2005
By Encyclopædia Britannica, Inc.

© 2005
By Encyclopædia Britannica, Inc.

Britannica, Encyclopædia Britannica, Macropædia, Micropædia, Propædia, and the thistle logo are registered trademarks of Encyclopædia Britannica, Inc.

Copyright under International Copyright Union
All rights reserved.

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

Printed in U.S.A.

Library of Congress Control Number: 2004110413
International Standard Book Number: 1-59339-236-2

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

CONTENTS

1 Decorative Arts and Furnishings
221 Delhi
226A Democracy
227 Denmark
244 The Great Depression
246 Diagnosis and Therapeutics
269 Dickens
275 Digestion and Digestive Systems
315 Dinosaurs
330 Diplomacy
341 Disease
394 Religious Doctrines and Dogmas
440 Dogs
451 Dostoyevsky
455 Drafting
460 Drawing
478 Dress and Adornment
529 Drugs and Drug Action
561 Dublin
566 Dutch Literature
569 The Earth: Its Properties, Composition, and Structure
615 The Earth Sciences
655 Earthquakes
667 East Asian Arts
772 Eastern Africa
838 Eastern Orthodoxy
857 Echinoderms
866 Eclipse, Occultation, and Transit
878 Economic Growth and Planning
908 Economic Systems
916 Economic Theory
953 Ecuador
963 Edinburgh
969 Edison

Decorative Arts and Furnishings

Decorative arts are those arts concerned with the design and decoration of objects, usually utilitarian, that in themselves do not necessarily possess aesthetic qualities. Certain of the decorative arts—basketry, pottery, and weaving, for example—are also collected under the term arts and crafts. By whatever name, they are distinguished from the fine arts in that the latter concern themselves with objects whose sole function is aesthetic appeal. The objects with which the decorative arts concern themselves range from humble household implements to monumental public works, and they include architectural units, furniture, rugs, and a host of items of metal, glass, clay, fabric, and other materials.

This article treats a number of traditional decorative arts in succession; the order in which they are discussed is suggested partly by historical development and partly by affinities. The opening section, an overview of interior design, sets a context in two senses: first, it begins with a general consideration of design that has implications for all the fine arts; and second, it discusses interior design as an art that makes extensive use of the products of the individual arts subsequently discussed in detail. Each section devoted to a particular form of decorative art includes discussions of materials and techniques and of significant historical developments.

The article is divided into the following sections:

Interior design 2
- Principles of interior design 2
  - Aesthetic components of design
  - Physical components of design
  - Design procedure
  - Kinds of interiors
- Historical developments 17
  - Origins of interior design
  - Interior design in the West
  - Interior design in the East

Furniture and accessory furnishings 44
- General considerations 44
  - Materials
  - Stylistic and decorative processes and techniques
  - Kinds of furniture
  - Kinds of accessory furnishings
- History 56
  - Western
  - Eastern

Rugs and carpets 69
- Elements of design 69
  - Field and border designs
  - Design execution
  - Colour
- Materials and technique 70
- Ornament and imagery 71
  - Individual motifs
  - Symbolism of overall design
- Uses of rugs and carpets 74
- Periods and centres of activity 74
  - Oriental carpets
  - Western carpets

Tapestry 81
- Materials 82
- Techniques 82
- Periods and centres of activity 84
  - Ancient Western world
  - Eastern Asia
  - Pre-Columbian Americas
  - Middle Ages in Egypt and the Near East
  - Early Middle Ages in western Europe
  - 14th century
  - 15th century
  - 16th century
  - 17th and 18th centuries
  - 19th and 20th centuries

Floral decoration 93
- Elements and principles of design 93
- Materials 94
- Techniques 95
- Forms of floral decoration 95
- Historical and stylistic developments 96
  - Western
  - Eastern
  - Other cultures

Pottery 101
- Kinds, processes, and techniques 101
  - Kinds of pottery
  - Forming processes and techniques
  - Decorating processes and techniques
  - Marking
- Western pottery 105
  - Ancient Near East and Egypt
  - Ancient Aegean and Greece
  - Etruscan and Roman
  - Islāmic
  - European: to the end of the 18th century
  - 19th century
  - 20th century
- East Asian and Southeast Asian pottery 126
  - China
  - Korea
  - Japan
  - Thailand and Annam
- American Indian pottery 136
  - North America
  - Central America
  - South America

Basketry 137
- Materials and techniques 138
- Uses 140
- Origins and centres of development 141

Metalwork 142
- General processes and techniques 142
- Western metalwork 143
  - Copper
  - Bronze and brass
  - Silver and gold
  - Pewter
  - Iron
  - Lead
- Non-Western metalwork 164
  - South Asia
  - Central and Southeast Asia
  - East Asia
  - American Indian peoples
  - African Negro peoples

Enamelwork 168
- Materials and techniques 168
- History 169
  - Ancient Western
  - Medieval
  - 15th century to the present: European
  - China
  - Japan

Lacquerwork 173
- Techniques 173
  - Obtaining and preparing lacquer
  - Application
  - Chinese carved lacquer
  - Japanese processes
- Historical development 174
  - China
  - Japan
  - Europe

Mosaic 179
- Principles of design 179
- Materials 179
  - Stone
  - Glass
  - Other materials
- Techniques 181
- Periods and centres of activity 182
  - Ancient Greek and Hellenistic mosaics
  - Roman mosaics
  - Early Christian mosaics

Byzantine mosaics
Medieval mosaics in western Europe
Renaissance to modern mosaics
Pre-Columbian mosaics
Stained glass 190
Elements and principles of design 190
Materials and techniques 191
Subject matter 193
Periods and centres of activity 194
12th century
13th century
Early 14th century
Late 14th, 15th, and 16th centuries
17th and 18th centuries
19th century
20th century
The history of glass design 199
Antiquity and the Middle Ages 200
Early glass
The Roman Empire
Byzantium
Islām
Mid-15th to mid-19th century 203
Venice and the façon de Venise
Germany
England
United States
Mid-19th to 20th century 210
Great Britain
United States
Czechoslovakia, Austria, and Germany
France
The Scandinavian countries
Belgium and The Netherlands
Italy
Chinese glass 214

INTERIOR DESIGN

Although man's desire to create a pleasant environment is as old as civilization itself, interior design, the conscious planning and design of man-made spaces, is a relatively new field.

Since at least the middle of the 20th century, the term interior decorator has been so loosely applied as to be nearly meaningless, with the result that other, more descriptive terms have come into use. The term interior design indicates a broader area of activity and at the same time suggests its status as a serious profession. In some European countries, where the profession is well established, it is known as interior architecture. Individuals who are concerned with the many elements that shape man-made environments have come to refer to the total field as environmental design.

The following section deals with principles of good design that are applicable to all design activities, with special emphasis on interior design, particularly as a creative and problem-solving activity.

Principles of interior design

It is important to emphasize that interior design is a specialized branch of architecture or environmental design; it is equally important to keep in mind that no specialized branch in any field would be very meaningful if practiced out of context. The best buildings and the best interiors are those in which there is no obvious disparity between the many elements that make up the totality. Among these elements are the structural aspects of a building, the site planning, the landscaping, the furniture, and the architectural graphics (signs), as well as the interior details. Indeed, there are many examples of distinguished buildings and interiors that were created and coordinated by one guiding hand.

Importance of total design

Because of the technological complexity of contemporary planning and building, it is no longer possible for a single architect or designer to be an expert in all the many aspects that make up a modern building. It is essential, however, that the many specialists who make up a team be able to communicate with each other and have sufficient basic knowledge to carry out their common goals. While the architect usually concerns himself with the overall design of buildings, the interior designer is concerned with the more intimately scaled aspects of design, the specific aesthetic, functional, and psychological questions involved, and the individual character of spaces.

Although interior design is still a developing profession without a clear definition of its limits, the field can be thought of in terms of two basic categories: residential and nonresidential. The latter is often called contract design because of the manner in which the designer receives his compensation (*i.e.*, a contractual fee arrangement), in contrast to the commission or percentage arrangement prevalent among residential interior decorators. Although the volume of business activity in the field of residential interiors continues to grow, there seems to be less need and less challenge for the professional designer, with the result that more and more of the qualified professionals are involved in nonresidential work.

The field of interior design already has a number of specialized areas. One of the newer areas is "space planning"—*i.e.*, the analysis of space needs, allocation of space, and the interrelation of functions within business firms. In addition to these preliminary considerations, such design firms are usually specialists in office design.

Specialized design firms

Many design firms have become specialized in such fields as the design of hotels, stores, industrial parks,or shopping centres. Others work primarily on large college or school projects, and still others may be specialists in the design of hospitals, clinics, and nursing homes. Design firms active in nonresidential work range from small groups of associates to organizations comprised of 50 to 100 employees. Most of the larger firms include architects, industrial designers, and graphic designers. In contrast, interior designers who undertake residential commissions are likely to work as individuals or possibly with two or three assistants. The size of the firms involved in nonresidential design is a clear indication of the relative complexity of the large commissions. In addition to being less complex, residential design is a different type of activity. The residential interior is usually a highly personal statement for both the owner and the designer, each of whom is involved with all aspects of the design; it is unlikely that a client who wished to engage the services of an interior designer for his home would be happy with an organized systems approach.

Most large architectural firms have established their own interior-design departments, and smaller firms have at least one specialist in the field. There are no precise boundaries to the profession of interior design nor, in fact, to any of the design professions. Furniture design, for example, is carried out by industrial designers and furniture designers as well as by architects and interior designers. As a rule, furniture designed for mass production is designed by industrial designers or furniture designers; the interior designer or architect usually designs those special pieces that are not readily available on the market or that must meet specific needs for a particular job. Those needs may be functional or aesthetic, and often a special chair or desk designed for a specific job will turn out to be so successful that the manufacturer will put such pieces into his regular line. The same basic situation holds generally true in the design of fabrics, lighting devices, floor covering, and all home-furnishing products. All design activities are basically similar, even though the training and education in the different design fields varies in emphasis. A talented and well-trained designer can easily move from one specialized area to another with little difficulty.

Distinction between art and design

In the discussion of the general aspects of design, it is important to note that there is an important distinction between art and design. A designer is basically concerned with the solution of problems (be they functional, aesthetic, or psychological) that are presented to him. The

artist is more concerned with emotive or expressive ideas and with the solution of problems he himself poses. A truly great or beautiful interior can indeed be called a work of art, but some would prefer to call such an interior a "great design."

AESTHETIC COMPONENTS OF DESIGN

A general definition of beauty and aesthetic excellence would be difficult, but fortunately there are a number of generally accepted principles that can be used to achieve an understanding of the aesthetic considerations in design. One must note, however, that such understanding requires exposure and learning; an appreciation of any form of art needs such a background.

A thorough appreciation of design must go beyond the first impression. The first impression of the interior of a Gothic cathedral might be that it is somewhat dark or gloomy, but, by the time the visitor senses its majestic proportions, notices its beautiful stained glass windows and the effect of light, and begins to understand the superb structural system that permitted builders of cathedrals to achieve their lofty goals, he can truly begin to appreciate the overall aesthetic qualities (Figure 1).

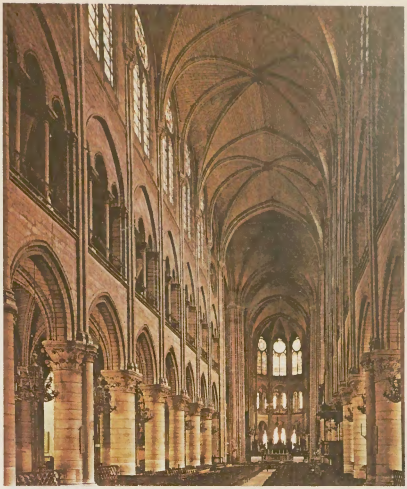


Figure 1: Majestic overall aesthetic quality of a Gothic interior: nave and choir, cathedral of Notre Dame, Paris, 1163–*c.* 1200.
Shostal—EB Inc.

Importance of function

One of the key considerations in any design must be the question of whether a design "works" or functions for its purpose. If a theatre has poor sight lines, poor acoustics, and insufficient means of entry and egress, it obviously does not work for its purpose, no matter how beautifully it might be decorated. Such a design could be considered good only if it were thought of abstractly as a kind of walk-in sculpture. In some cases the building is meant to be sculpture rather than architecture. The Statue of Liberty, for instance, is primarily intended as a monument, despite the fact that it contains rather tortured interior spaces.

To use function as the only aesthetic criterion would be limiting, but it certainly is a valid consideration to be kept in mind. Designers are often tempted to overdesign or "style" an object or interior rather than design it. Some of the most beautiful objects of the 20th century are beautiful because they were the result of purely functional considerations. It is conceivable that future art historians will consider a modern jet plane the crowning artistic achievement of the middle of this century, rather than any building, interior, or conscious art form.

The aesthetic response to an interior and its furnishings must take into consideration the social and economic conditions as well as the materials and technology of the time. The elegant or ornate interiors that are usually associated with the 18th and 19th centuries were appropriate to the social and economic conditions of the nobility or the wealthy bourgeois who were the original occupants. The chairs were designed for formal living, and the elaborated carved furnishings were designed to be cared for by many servants (Figure 2, left). Such an interior is alien to the 20th-century way of life and would be totally inappropriate for a contemporary middle class family. It would also be inappropriate to use modern materials and processes to imitate earlier materials and processes (Figure 2, right). Many manufacturers try desperately to make plastic look like wood, stone, or just about anything but plastic. All

By courtesy of (right) Knoll International, Inc.; photograph, (left) Louis Reens





Figure 2: *Social and economic considerations in interior design.*
(Left) Elaborate mid-19th-century dining room in the Gothic Revival style, Lyndhurst, Tarrytown, New York, designed by Alexander J. Davis. (Right) Simple pedestal table and chairs appropriate to the dining room of the mid-20th century family, designed by Eero Saarinen, 1956–59. Synthetic materials and mass production methods are used to achieve furniture suited to its function.

aesthetic criteria have something to do with honesty. Some aesthetics have compared beauty to truth, and there can be little doubt that honestly expressed functions and honestly expressed materials and manufacturing processes are far more beautiful than fakery and imitation.

Relation of interior to basic structure

All interiors, by definition, occur inside buildings and therefore have a very real relation to these buildings. The best interiors today, as well as in the past, are those that relate well in character and appropriateness to the particular building. The furnishings designed and scaled for spacious country homes or palaces would obviously be out of place in a small urban apartment or suburban home. A strong and unusual piece of architecture such as New York City's Trans World Airlines terminal (at John F. Kennedy International Airport) could not be properly furnished with standard commercial furniture and products. The building, as well as the interiors, was conceived as a total design by the Finnish-born architect Eero Saarinen. Whether the observer agrees with the architect's concept or not, he clearly senses the strong interrelationship between the exterior and the interior—and therefore the aesthetic unity and success. Another successful interior and building is the Ford Foundation headquarters in New York City, the work of architects Kevin Roche and John Dinkeloo, with interiors by Warren Platner. The design is notable for its handsome spaces opening out toward an enclosed garden space (Figure 3). This obviously would not have been possible or appropriate if the view from the offices had been unattractive.

J. Zimmerman—FPG



Figure 3: Aesthetic unity in the interrelation of exterior and interior space: Ford Foundation headquarters, New York City, designed by Kevin Roche and John Dinkeloo, 1967.

The interiors within indifferent or unattractive buildings must strive to make up for the lack of design qualities in the structures. Thus, it is sometimes necessary to ignore the ugliness of the building and create an inward-looking beauty if no architectural character exists.

The most difficult aesthetic consideration is the problem of appropriateness. The appropriate atmosphere or character of an interior must take all the foregoing points into consideration. The architectural character of the TWA terminal would make it inappropriate for use as an office building. The appropriateness of individual, more intimate, and small-scaled interiors is more subtle. The interior design of a discotheque would hardly be appropriate for a research library, and a college classroom would hardly provide the desired atmosphere for a kindergarten. Many of these responses and relationships are complex and have psychological as well as aesthetic factors.

Elements of design. Of all the component elements that together form a completed interior, the single most important element is space. Spaces can be exhilarating or depressing, cheerful or serene, all depending upon the use the designer has made of the various elements that form the whole. Space is, in modern times, a costly commodity. The beautiful space of the Gothic cathedral achieved its success through generous proportions and lofty heights (see Figure 1). Due to the vast increase in construction costs in contemporary structures, spaces tend to be smaller and less generous; more skill on the part of the designer is required to give such limited spaces a particular atmosphere or character. On the other hand, sheer volume of space is not sufficient. There is hardly a larger space than the interior of the Vehicle Assembly Building at the John F. Kennedy Space Center in Florida, yet the aesthetic impact of that immense interior is negligible. A space need not be large and monumental to be aesthetically successful. The handling of mass and form even within a small structure can become exciting and beautiful. Frank Lloyd Wright was masterful in creating beautiful spatial sequences within residential-scale buildings (Figure 4). The Ford Foundation building is a relatively small structure among the huge buildings of New York City, yet the experience of that space is real and pleasurable.

Characteristics of space

Space can be thought of as the raw material which must be molded and shaped with the designers' tools of colour, texture, light, and scale. The interrelationship of design elements can be clarified by visualizing the result if the interior of St. Peter's in Rome were painted in garish colours or painted all black or sprayed with a foamy texture covering all surfaces or flooded with enormously intense floodlight that eliminated all play of dark and light. Obviously, any of these modifications would totally destroy the beauty and success of that space.

Colour is the quality of light reflected from an object to the human eye. When light falls upon an object, some of it is absorbed, and that which is not absorbed is reflected, and the apparent colour of an object depends upon the wavelength of the light that it reflects. The scientific attributes of colour and light in interior design are, however, less important than the skillful combination of colour values, hues, tones, shades, and above all textures. Although there can be no strict rules about colours and textures, it is well to remember the famous statement of the modern architect Mies van der Rohe that "less is more." His Crown Hall at Illinois Institute of Technology in Chicago, built in 1956, is elegant, understated, subtle, and is notable for its careful handling of textures and materials. To accept "less is more" as the sole guideline to design, however, would be a serious fallacy. Space, which is the essence of a meaningful interior, would be dull indeed if it were never varied—if there were no intimate spaces with low ceilings, in contrast to large spaces of greater height, and if spaces did not interrelate to provide the user with a sequential experience of moving from one to another. Monotony would also result if all interiors in a given building were of the same colour, material, and textural quality. Man needs variety and change.

The use of colour

The manipulation of space is a matter of both aesthetic and functional consideration. A small entrance vestibule in a building is needed to keep out wind and cold or heat and rain, yet it is equally important in providing a visual transition from outdoors to the interior of the building. The sheltered sleeping alcoves in early cave dwellings served not only to express man's desire for smaller and more intimate spaces for personal use but gave protection from draft or cold.

Much in our man-made structures is built of natural materials, and it must be remembered that these materials have natural colours and textures that usually are supe-

Value of natural properties



Figure 4: Carefully modulated spatial sequences in residential scale exemplified by the living room designed by Frank Lloyd Wright for his home and studio, Taliesin East, at Spring Green, Wisconsin; photograph, 1939.
Hedrich-Blessing, photo

rior to anything man can create artificially. Competent designers are very much aware of the innate qualities and textures of all materials, especially natural ones (see Figure 4). For instance, a sensitive designer would choose a simple oil finish on wood to bring out the beauty and quality of the grain rather than use the once-fashionable high-gloss finish that tended to obscure and change the texture. Textures are important not only for their appearance but also for their sense of touch, and for their effect on light absorption or reflection. Abrasive surfaces or very rough plaster would obviously be unpleasant to the touch and possibly dangerous in an interior, depending upon the use the interior is intended for. Textures can evoke feelings of elegance (such as silks) or informality (such as rough, tweedy materials).

Light, both natural and artificial, is one of the most important design elements, but unless surfaces are appropriate in colour and texture, the control and effect of light will be lost. The beautiful quality of space in a Gothic cathedral is very much related to the handling of light (see Figure 1). The source of daylight, high overhead or filtered through stained glass, creates exciting patterns of light and shade and a variety of intensities and pools of light. This same principle can be used in all interior spaces, and contemporary interiors often have skylights or high windows to provide variety and changing patterns of light. Artificial lighting is equally important, and, again, the same considerations of highlights, good overall illumination, and variety are important.

Concepts of design. The scale and proportion of any interior must always relate to the architecture within which the interior exists, but the other important factor in considering the scale of man's environment is the human body. Throughout the ages, designers and architects have attempted to establish ideal proportions. The most famous of all axioms about proportion was the golden section, established by the ancient Greeks. According to this axiom, a line should be divided into two unequal parts, of which the first is to the second as the second is to the whole. Leonardo da Vinci developed a figure for the ideal man based on man's navel as the centre of a circle enclosing man with outstretched arms. The French architect Le Corbusier developed a theory of proportion called Modulor, also based on a study of human proportions. Yet, at best, these rules are merely guidelines. They can never substitute for the eye and judgment of the designer, and it is reasonable to predict that attempts to make the all-powerful computer a substitute for the designer's sensitivity are also bound to be far from perfect.

The search for ideal proportions

It was stated earlier that the need for a changing scale and spatial relationship in the environment seems a natural one, almost a physiological as well as a psychological one. Perhaps the need for "personal" environment and scale can best be understood by considering some extreme examples. To a person flying at 30,000 feet in an airplane, the scale of anything seen on the ground appears so small that he loses touch with the reality of objects. People who fear heights are rarely bothered by the view out of an airplane because the distance to the objects on the ground has transcended normal perceptions of scale. In a similar manner, a person's reaction to the scale of a small house is quite different from his reaction to a large high-rise building. Details of pattern, texture, and material are accepted and expected in the small structure since they are in a meaningful scale with respect to man. By the same token, the sculptural ornaments on the tops of early skyscrapers seem absurd today.

Almost all principles of design for interiors can be comprehended with clear analytic understanding and common sense, without regard to dogmatic rules. If a beautiful 18th-century breakfront (which might be more than eight feet tall) is placed in an apartment with a ceiling height just an inch higher than the piece of furniture, it would obviously look out of scale. If a space is planned so that all the heavy and massive pieces of furniture are pushed toward one end of the room, with nothing on the other side, the room would obviously look out of balance. Yet balance and symmetry applied as inviolate design principles would result in very formal, very traditional, and somewhat dull interiors. Careful symmetry was a generally accepted rule during the Renaissance, and in any classic building one can be sure to find a carefully balanced and symmetrical facade, just as most formal and classic interiors have rigidly balanced plans. It is now recognized that balance can also be based on asymmetry. Both architecture and interior design in the 20th century have consciously broken with the many rules handed down from past eras. It is more important for a building or space to be expressive of its purpose. At one time, it was traditional for a theatre, opera house, or concert hall to embody certain forms and shapes without any real consideration of sight lines, seating distance from the stage, or acoustics. On the other hand, the Berlin Philharmonic Concert Hall (1964) works beautifully as a concert hall and expresses its purpose and function clearly in an exciting and dynamic way (Figure 5).

Departures from the rules in the 20th century

Balance and symmetry, colour, pattern, and repetition used to be a matter of adherence to a tradition. Until fairly recently, many interiors were painted in dark colours, often ignoring the fact that light reflection was adversely affected and that no real contrast or sparkling accent was achieved. In many contemporary rooms, however, most surfaces are kept in neutral or light colours, possibly with one wall accented in a strong colour or texture. An interior with uniform overhead lighting might be an efficient work space but would lack the character that can be achieved by providing some accent lights in small areas.

The designer's concern for honesty of materials and textures has brought about changing attitudes toward some of the conventional practices of interior decoration, such as the use of strongly patterned wallpapers and flowered prints. Any interior that has too many different patterns, too many textures, and too many repetitive features of any kind will appear overpowering, overly busy, overdesigned, and confusing. A designer often attempts to have a dominant theme or idea, be it colour, form, texture, or some rhythmic pattern. It must be noted also that design is influenced by changing attitudes and fashions. The movements in art and architecture of the 1950s and 1960s have influenced interior design in the direction of an emphasis on pure form, the absence of superfluous decoration, and expressiveness of materials. Recently, however, a kind of countermovement in the field of painting and sculpture has been influential. For instance, the use of large-scale graphic elements (supergraphics) in interiors has become popular and accepted, in spite of the fact that its very idea often consciously denies or destroys the visual clarity of existing architectural design features. Some of the leading designers in the United States and in several European

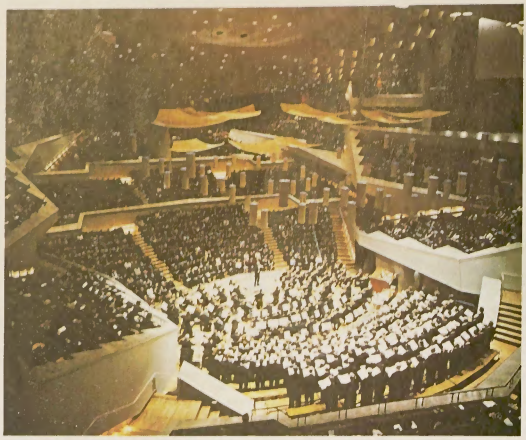


Figure 5. Dynamic, asymmetrical architecture creating an unconventional yet functional interior design space. Berlin Philharmonic Concert Hall, designed by Hans Scharoun, 1964.

By courtesy of the Staatsbibliothek Preussischer Kulturbesitz Bildarchiv, Berlin

countries have also become very interested in large patterns, rhythmic geometries, and decorative surfaces, and this may point toward a new trend (Figure 6).

Sequential relationships

Most interiors consist of a series of interrelated spaces. It is important that the various spaces be designed in a sequential relationship to each other, not only in terms of planning but also in terms of the visual effect. A successful interior should be cohesive within each area and cohesive as a totality. It must above all relate to the building and to the architectural concept. A good example is the previously mentioned TWA terminal by Eero Saarinen. In spite of the extremely complex sculptural forms used, there is a sequence and clearly balanced rhythm that not only unifies the total composition but clearly relates it to the total architecture.

Jeremiah O. Bragstad



Figure 6. Supergraphic interior emphasizing decorative rather than architectural design. Hear-Hear Record Shop, San Francisco, designed by Daniel Solomon, graphics designed by Barbara Stauffacher, 1969.

The best examples of design are those in which no visible difference exists between the interior and the exterior, between the building and its site, and between the many parts or spaces to each other and the total building. An example is the house of the American architect Philip Johnson in New Canaan, Connecticut. Johnson's home and its setting appear effortlessly united, with individual parts subordinated to the success of the whole (Figure 7).

Design relationships. The real and conscious relationship between art, architecture, and design is of long standing. Though mural painting was largely neglected in the mid-20th century, in the past great murals have been the planned focal points of interiors and have in a way determined the architecture (Figure 8). Similarly, sculpture or sculptural forms, as fixed and permanent spects of buildings, can be the most important design features if planned that way by the architect together with the interior designer and artist. Perhaps the best design is one in which there is no visible difference between architecture and interior and in which even the artwork is incorporated as an integral part of the total (see Figure 14).

The design relationship of interiors to architecture can be clarified by citing an extreme example: the stage set. A set for a theatrical production is a form of interior design but, unlike all other aspects of interior design, it attempts to create its own world and atmosphere concerned only with the play and not at all related to the world or even reality. The creation of a world of make-believe is precisely the function of a stage, but in real life it is impossible to divorce a particular interior from everything else around it. Sometimes a designer may attempt to create a "theatrical" interior, but the point being made strongly and unequivocally here is that every interior must relate to the architecture and to the nearby environment.

Art objects and interior design

Design relationships of individual works of art (paintings, prints, or sculptures) to interiors are most significant in terms of scale and placement, rather than in terms of subject matter, colour, or style. A very old painting, if it is good, will look well within a contemporary interior; a very modern piece of sculpture can be beautiful within an interior furnished with some beautiful traditional pieces. Any work of art, if successful within itself, is "correct" with any interior if properly placed or selected to work with the total space. Certainly there is no need to match





Figure 7: *Interrelation of interior and exterior space.*
Harmony of landscape, architecture, and interior design: Glass House, New Canaan, Connecticut, designed by Philip Johnson, 1949. (Left) Exterior. (Right) Interior.
Russ Kinne—Photo Researchers

colours of paintings to interiors or to select subject matter in works of art that reflect a particular theme, such as food for dining rooms or hunting scenes for the den.

Interiors as they relate to landscape or cityscape are sometimes misunderstood by architects. A crass but typical example is the ubiquitous picture window in suburban housing tracts. Often the only view from the window is the picture window of the neighbouring house. When the view is a beautiful one, it should be possible to plan the interior with the furniture plan and orientation such that seating arrangements can take advantage of the view and yet work for other functions, such as relation to a fireplace or a conversation group, as well.



SCALA — Art Resource

Figure 8: A simply designed interior space made vivid and compelling by frescoes on the ceiling and walls: Sistine Chapel, Rome, by Michelangelo, 1508–12, 1533–41.

Graphics in interior design

In many areas of interior design the field of graphics is taking on increasing importance. In every public or institutional building, signs, directories, and room identifications play an important visual part. Good architectural graphics have been stressed only in recent years, as a result of the increasing size and complexity of structures. Buildings such as airports depend upon clear and handsome graphics to make the spaces work and to make them aesthetically cohesive. A related aspect of graphics is the printed matter that is part of certain interior functions. Interior designers must be concerned with the design of menus, wine lists, napkins, and matchbooks in a well-designed restaurant. Designers dealing with stores or shops are concerned with the graphics of shopping bags, signs, and posters. Often the interior designer is the actual graphic designer, or he works together with the graphic designer, just as the architect works with the interior designer or landscape architect.

Modes of composition. It must be emphasized that there are many different moods, or modes of composition, that are possible in interior design. The recognition of this fact makes it difficult to apply valid critical criteria to these modes, since many of them are intensely personal. What may appear to be picturesque to one person might be ugly or cluttered to another. Each person brings to interior design his own cultural mores and his own prejudices, and in many ways he is psychologically conditioned and influenced to accept certain things and to reject others. In discussing various modes of composition, one must therefore take into consideration the occupants and their backgrounds, the locale and site, and then try to apply the most basic design principles as general guidelines.

Formal and informal compositions

Formal and informal compositions are relatively easily defined and classified; in fact, this distinction is useful throughout the history of furniture and interiors. Formal styles are usually associated with life at court or furnishings for the palatial homes of nobles or a moneyed elite. The informal periods usually are associated with rural living or the simpler pieces of furniture made by the local craftsmen in rural areas, where they plied their trade with limited tools, using local woods. Formal furniture, as a rule, leads to formal interior compositions. Balance and symmetry certainly tend to lead to formal compositions. Formality is not associated with any particular period. In fact, a very famous contemporary chair, the Barcelona chair by Mies van der Rohe (Figure 58), is an extremely formal and elegant piece. It would seem wrong to use that chair in a casual catercorner room arrangement.

Setting strongly influences the character of a space. By its very definition, a rustic setting would be rural and infor-

mal and would seem wrong and incongruous in a formal townhouse or city apartment. Since most business and public interiors are located in urban centres, any attempt to make such interiors look rustic or homey would be an aesthetic paradox. By the same token, it would appear equally incongruous to design a restaurant located in an old mill or barn in New England in a formal and urban character with elegant furnishings, whether they were contemporary or antiques of a formal nature.

Functional compositions

Certain modes of composition are determined by function of the spaces as much as by the location and by the architecture. For example, a cozy or homey interior is normally associated with residential interiors or similarly intimate interiors, such as restaurants that may wish to appear "cozy." Some interiors, such as discotheques, require excitement and other interiors, such as funeral parlors, require serenity or dignity. One expects certain modes of composition for certain functions, but one's expectations are subject to many external influences, such as personal background, location, psychological associations, and changing fashions. For instance, the typical bank interior until about 1950 was expected to be solid, dignified, awe-inspiring, formal, and above all confidence inspiring. Contemporary design for business and industry has become accepted by all, and the early 1950s saw the logical extension of these firmly established design principles into the area of bank design. One of the first radical departures of traditional design for banking spaces was the Manufacturers Trust Company Manhattan office designed by Skidmore, Owings and Merrill in the early 1950s. It was the first widely published "glass" bank, and it set a trend that has become the new mode of composition for banks.

Fashion or design trends influence one's reactions to many kinds of designs. The term clutter is usually associated with Victorian design of the 19th century. Under the usual definition of the term clutter, one thinks of home interiors with collections of accessories and with an overabundance of knickknacks—the typical Victorian home (Figure 9). In the mid-1960s a new approach to office design, reflecting the "cluttered" approach, was developed. This office appears disorganized at first glance. Actually, the system (called office landscape; see below *Kinds of interiors: Public interiors: Space planning*) is very efficient and for that reason is deemed acceptable, even if the visual impact tends to be chaotic. Traditionally, office and business interiors were pristine, orderly, and very organized, and the idea of a cluttered appearance would have been anathema to designers.

Exotic compositions

The most difficult mode of composition for objective analysis is one that some people call exotic. The chances are that all exotic interiors are highly personal statements and cannot be rationally understood in theoretical design terms (Figure 9). To begin with, what may appear exotic to the average American could be very ordinary or even homey to another culture. Japanese or oriental design in general serves as an example. A Japanese style interior is extremely subtle, serene, and understated, yet to the uninitiated such an interior will appear exotic. Undoubtedly that same phenomenon holds true in reverse. Oriental people have often been impressed with Western-style design and have adopted it presumably because to them it appeared exotic. The increased mobility of the middle classes of many nations today has made foreign travel possible for more and more people, thereby tending to soften some of the very strong regional differences in design. The modes of composition are still discernible nationally or certainly by major geographic and ethnic divisions, but they tend to be less distinct. Many subtle differences exist within the same country, some of which are based on varying socioeconomic backgrounds, much in the manner of the traditional difference between formal styles (at court and in homes of nobility) and informal modes of composition for the country people and middle classes. The labels that one applies to these modes of composition are often only descriptive. They must not be confused with objective evaluation of design values. An interior that is by the creator's definition exotic or picturesque may or may not be a well-done exotic design.

Symbolism and style. There are many historic examples

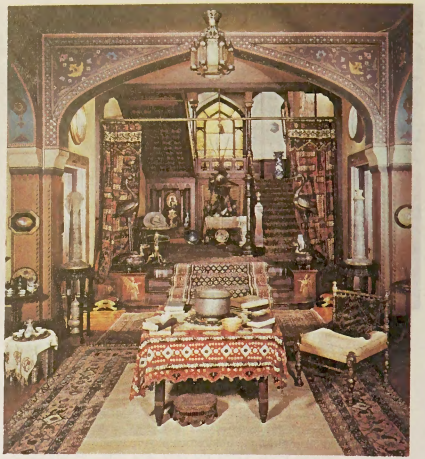


Figure 9: A cluttered Victorian interior in the exotic Moorish style, designed by the landscape painter Frederick Edwin Church for his home, Olana, at Hudson, New York, 1870–72.
Frank Lerner

of symbolism in design, but often the symbolism is not a conscious statement so much as a more subtle reflection of style. Religious buildings, especially churches, have until recently been consistently traditional expressions of style or symbolism. The church and church architecture flourished during the Middle Ages, and the style of church architecture that became the dominant symbol was the Gothic style. Until the recent past, churches were still designed, as a matter of course, in Gothic style. It is interesting to note that a "Gothic" church designed and built

Symbolism in religious buildings

By courtesy of United Airlines, Inc.



Figure 10: An interior shaped by objects symbolizing Theodore Roosevelt's personal interests and personality, North Room, Sagamore Hill, Oyster Bay, Long Island, 1880.

in 1820 can be clearly identified as such, and a "Gothic" church from the year 1920 has the imprint of that year as obviously as the date on its cornerstone. There has been a similar symbolic or stylistic tradition in the design of public or governmental buildings. Both interiors and exteriors of city halls, court buildings, and major government structures were usually in the "classical" style, symbolizing authority, power, and stability, based on our long historic association of these concepts with Greco-Roman antiquity and Renaissance thought.

Another form of symbolism in interior design has been the creation of interiors around specific themes or concepts. Among the earliest examples is the Egyptian tomb. The interior design and decoration depicted the life of the king or special events from his life, and the total interior was intended as a kind of magic to assure the occupant's journey into life after death and guarantee his happiness there. Another example of a symbolic interior created for a specific purpose is the Roman hunting lodge, Piazza Amerina, in Sicily, which has splendid murals and floors depicting animals and hunting. A more recent example of a similarly symbolic interior on the same subject is Theodore Roosevelt's home at Oyster Bay on Long Island, built in 1880. It is full of hunting trophies and mementos symbolizing his personal interests and his personality (Figure 10).

The styles that developed in interiors and in interior furnishings were always symbolic of the social structure of the society that created them. It is easy, for instance, to look at the graceful, feminine lines of a Louis XV chair, delicately curved and luxuriously upholstered, and to see it as a symbolic expression of the superficialities of court life. One can also look at some of the crudely fashioned early American furniture and see in one's mind the life of the settler who fashioned it. Life was harsh, time was precious, and articles of furniture were confined to essentials. The need for economical use of space was symbolized by dual-purpose, functional pieces such as dough boxes that served as tables and tables that turned into chairs and had storage compartments for the family Bible as well.

Symbolism in business offices

As functional and efficiency-oriented as business and office design is today, it is full of unwritten rules relating to symbolism. The design of an office reflects the status of the occupant. Top executives are located in the largest corner offices with the best views of the city and invariably are on the top floors of the corporate headquarters. The size of desks is a symbolic indication of the executive's importance in the hierarchy of the firm. The very top officers may, however, do away with desks altogether and have offices resembling living rooms—to symbolize the fact that they are beyond routine paperwork and above the need for standard office furnishings. The fashions (or styles) of design vary and develop even within a brief period of 10 or 20 years. Thus, another symbol—carpeting—has become somewhat outdated. Until recently, top executives expected wall-to-wall carpeting in their offices. Today such offices may have wood or other natural floors, perhaps with beautiful area rugs. The very idea of a private office is, of course, the most important symbol in a status-conscious business community (Figure 11). Designers have found, however, that the need for communication between executive and staff, including visual contact, often makes private offices less than efficient.

Symbolism in residential interior design occurs on many levels but again tends to be influenced by changing styles. When television first became available, the home screen became a symbol of prosperity and at the same time became the focal point of residential interiors. By the 1970s a television set had become a standard possession and was no longer a compositional emphasis; in fact, it was often concealed or casually incorporated into the total design.

A homeowner is likely to be very conscious of the image his house or apartment conveys. Traditional furniture, for instance, is still associated with elegance in the minds of many laymen, a situation that can lead to the acquisition of poor reproductions or meaningless imitations of nonexistent styles. To most people a real fire in a fireplace is a delightful physical and visual experience that often has nostalgic associations. Since they are no longer needed to heat houses, fireplaces in the 20th century increasingly have become a luxury and thereby a symbol of substance to many people. These circumstances have often resulted in imitation fireplaces of the worst possible design, with simulated fires.



Figure 11: Executive office resembling a residential interior: Fabergé Corporation Headquarters, New York City, designed by Dallek Inc., Design Group, 1968.

By courtesy of Dallek Inc., Design Group

From the designer's point of view, design symbolism in public spaces is valid at times but can and should be used in contemporary terms rather than as stylistic imitation of past eras. An example of the success of such design can be seen in the new Boston City Hall, built in 1968, which symbolizes government, authority, and dignity in totally original and contemporary terms. There is little valid reason to consciously introduce symbolism into residential interiors, unless it is the kind of cultural symbolism exemplified in Japanese interiors, such as that of the Zen tea house (*cha-shitsu*), where certain design features reflect a way of life and have ceremonial meanings.

Contemporary symbolism

PHYSICAL COMPONENTS OF DESIGN

The foregoing section on aesthetic components stressed the fact that, in design, the whole or total effect is more important than the specific device or element used. The same is true of architectural components, and this should be kept in mind in the following discussion.

Ceilings. Although ceilings are in most interiors the largest unbroken surface, they are often ignored by amateur designers and even by professional designers. The result, especially in public and office interiors, is frequently a mass of unrelated lighting devices, air conditioning outlets, and the like. Ceilings were emphasized in the Baroque and 18th-century traditions: beautiful interiors of these periods had highly ornate, decorated ceilings, with painted surfaces or with intricate plaster details and traceries (Figure 12, left).

Few modern designers take advantage of the design possibilities offered by ceilings. One such possibility is the creation of textural effects with wood. Of course, one must respect the effect of a simple plaster ceiling in an otherwise well-designed interior; often the white plaster ceiling is needed to reflect light and to provide a calm cohesiveness to the space (Figure 12, right). Since most modern ceilings are low, a heavy texture or a strong colour could create a depressing feeling; hence, the popularity of a plain white ceiling. It is important for a plain ceiling to be just that: a surface without blemishes, without bumps, and without small unrelated areas of different height.

Value of the plain white ceiling

In contemporary public buildings there is frequently a "hung" ceiling below interior concrete structural slabs.





Figure 12: *Ceiling design.*
(Left) Highly ornate Rococo ceiling, Pilgrimage Church at Wies, Upper Bavaria (Germany), designed by Dominkus Zimmermann, 1745. (Right) Simple, white plaster ceiling, Christ Lutheran Church, Minneapolis, Minnesota, designed by Eliel and Eero Saarinen, 1950.
(Left) Toni Schneiders, (right) Balthazar Korab

The space between the slab and the "hung" ceiling is needed for mechanical equipment as well as to allow for the recessing of the lighting system.

An earlier section of this article discussed the variation of heights in relation to scale and space. It is important to keep such varying ceiling heights related to the plan of the room if such a device is to succeed. A lowered ceiling in a dining area, for instance, can be pleasant and intimate, but a lowered ceiling covering only part of the area can be most distracting.

Floors. Basically, there are two kinds of floors for interiors: those that are an integral part of the structure and those that are applied after the structure is completed. Interior designers working together with architects have the opportunity to specify flooring such as slate, terrazzo, stone, brick, concrete, or wood, but in most interiors the flooring is designed at a later stage and is often changed in the course of a building's life. Sometimes it is possible to introduce a heavy floor, such as terrazzo or stone, in a finished building or during remodeling, but these materials, beautiful as they are, tend to be too costly as surface applications.

Resilient flooring materials

Man-made, or synthetic, floor coverings are usually classified as resilient floors. The oldest of this type is linoleum. The resilient flooring materials marketed in the late 20th century include asphalt, vinyl asbestos, linoleum, cork, and vinyl. Cork, which is not a synthetic, is handsome, but is difficult to maintain and is not exceptionally durable. Basically, other resilient floor tiles are excellent flooring materials that are both economical and easily maintained. They can be given almost any appearance, which is a temptation that manufacturers are unable to resist. When the tiles are plain, in good colours or textures, they are very attractive and appropriate, but often they are made to imitate stone, brick, mosaic, or other materials, and the results are generally of a less satisfactory nature. Pure vinyls are the most expensive of the resilient floorings and have been the most tortured in terms of "design." The vinyls are the softest and most resilient of the tiles and are very easy to maintain. Asphalt tile is the least expensive and consequently the most widely used resilient flooring, although it is quite brittle and hard underfoot. Vinyl asbestos is somewhat softer underfoot and, being grease resistant, is easier to maintain than asphalt, but its cost is generally higher. Linoleum, which ranges in cost between the asphalt and pure vinyl floorings, is strong and suitable for heavy-duty uses.

Ceramic tiles and quarry (unglazed) tiles are made not only for such areas as bathrooms but, particularly in the case of quarry tiles, are suitable for almost any space. Installation usually requires a cement bed over the existing subfloor, making this material difficult to use in existing buildings. Like other natural materials, quarry-tile floors possess a natural beauty and have the additional advantage of easy maintenance.

Wood floors still account for a very large percentage of all floors, especially in residences. In addition to the strip oak floors, the standard for many apartment houses or homes, many beautiful prefabricated parquet patterns are available in a variety of woods and in many shapes and sizes. These wood tiles can be installed, just like the resilient floor tiles, over existing floors. Wood floors have great warmth and beauty but have the disadvantage of needing more care than do some of the synthetic tiles or quarry tiles.

Walls. Every wall is a material in itself; and ideally no material, if it is properly used, needs to be covered up. Some elegant buildings constructed since 1960 have used concrete in its natural texture—*i.e.*, showing the formwork left by wooden forms as a conscious expression of the material. During the 19th century, fakery in design was very popular, and part of the concern with the true expression of materials today is a revolt against the earlier tradition. In the 20th century, for instance, interior brick walls are considered very beautiful and desirable, yet many old townhouses have layers of plaster and paint or wallpaper on top of attractive brickwork.

It is not unusual for a decorative detail or device to survive long after the valid reason for it has disappeared. Wall panelling has been popular for hundreds of years, and, indeed, a natural wood texture adds warmth and elegance. The only way the craftsmen of earlier periods were able to apply wood panelling was in frames (stiles and rails) or wainscotting, since wood panelling was made of solid wood and had to be broken up into narrow dimensions in order to prevent warping and shrinking. Out of that need developed beautiful details of moldings, carved details, and carefully proportioned panelling. A similar art devel-

oped somewhat later in plaster. Obviously, 20th-century building costs and methods rarely permit real quality in elaborate panelling or highly ornate plasterwork (Figure 13), nor would this sort of imitative design be appropriate in a modern building. But wood panelling and plywoods in many beautiful veneers are readily available and provide a vast range of beautiful, if expensive, wall surfacing for important spaces. Prescored, pre-finished inexpensive plywoods, on the other hand, are often used as finishing materials for basement, recreation, or utility rooms in many homes in the United States.

Walls of wood veneer

The use of fake moldings, with printed moldings or panelling or with any of the countless imitation wall-surfacing materials from brick wallpaper to artistically poor wall murals, is the kind of decoration that a good designer avoids. Even so, not every interior should be a plain space with nothing but the natural walls. Highly decorative wallpapers have long been available in bold and exciting patterns. Often in 20th-century design a strong paper is employed on one wall only, instead of having the whole space surrounded by a dominant pattern. Many wallpapers, such as grasscloth and shiki silk papers from the Far East, have natural textures. For public spaces and for any space requiring easy maintenance and special cleanliness, a number of wallpapers have been developed that are completely washable and sanitary. Most of these are vinyl-coated fabrics, and some of them are extremely strong and durable and are particularly suited for such spaces as hospital or hotel corridors. Because these vinyl-coated wall fabrics are usually specified by designers and architects, the level of design is far superior to those made for the home.

There are many wall-surfacing materials using fabrics laminated to paper. These coverings provide warmth and texture, as well as acoustic properties. Fabrics in general have been used widely as wall-coverings in the past and continue to be popular.

A designer's imagination and the client's budget are the only limitation on the materials that may be used for wall surfacing. Some, such as ceramic or mosaic tiles, are extremely practical; some, such as cork, have excellent acoustical characteristics. For functional or for aesthetic reasons the designer may elect to use such materials as leather, metals, plastic laminates, or glass. No wall in itself should be designed or selected without relation to the total scheme.

Windows and doors. Windows and doors in contemporary design are not placed as decorative elements or as parts of symmetrical compositions but are primarily considered as functional elements and are expressed as such. If windows are carefully designed and placed for light, for ventilation, for air, and for view, decorative treatment is often unnecessary and a simple device such as a shade or shutter will suffice to control light and privacy. Most buildings, however, need window treatments, since no particular care in the placement of fenestration was taken by the builders.

Curtain characteristics

The most frequently used devices are curtains and draperies. Although semantically there is no clear distinction between the two, drapery implies more elaborate treatments with lining, overdrapes, valances, and tassels. A curtain, on the other hand, is lighter, more direct, less theatrical, and more functional. Frequently, a light material is chosen to provide privacy or light control with minimum emphasis. Curtains, however, offer only partial control over light, glare, and privacy; complete control or privacy often requires shades, blinds, or shutters. Window shades without overly ornate borders and tassels are a perfectly good device for those controls, and Venetian blinds are also a most acceptable treatment.

Since the 1960s designers have tried to simplify window treatments, and, if curtains, shades, or blinds were not deemed appropriate for functional or aesthetic reasons, devices such as chains or beads on windows or very simple sliding panels were found to be more effective than more elaborate treatments.

The essential considerations for windows must be based on the functional needs and on the overall aesthetic intent. If a space is well designed in architectural terms and presents a cohesive image, it rarely makes sense to feature a window or door. Poorly detailed windows in office buildings or apartment houses are often overcome or played down by using a simple curtain material covering a complete window wall. The wall-to-wall and floor-to-ceiling treatment of a window wall is frequently the only way to screen out unattractive details.

Planning for doors

Doors must be carefully planned, relating the swing and location to the functional needs, and their heights, colour, material, or textures to the adjoining wall surfaces or design elements in the space. Most doors used in the 20th century are "flush" doors—that is, they have unbroken surfaces made of wood or metal; even where glass is used the attempt is usually made to have maximum glass area unbroken by frames and moldings. Sometimes the entrance doors to important spaces are designed or decorated as compositional focal points, but usually the emphasis is on excellence in detailing and hardware rather than on decorative surface designs.

By courtesy of the Metropolitan Museum of Art, New York. Fletcher Fund, 1931



Figure 13: Ornate plasterwork to decorate wall and ceiling: dining room from Kirtlington Park, Oxfordshire, designed by Thomas Roberts, completed 1748. In the Metropolitan Museum of Art, New York City.

Other components. The detailing referred to in connection with the handling of doors is one of the most important factors in interior design. Every architectural component must be detailed well. Poor details make for poor design. The meaning of detailing in a design sense is more than the graphic explanation of certain components on a drawing. It means the way materials are put together, the way one part is fastened to another, the way parts and materials are expressed and articulated. Stairs or ramps are architectural components of great importance, whether in stores, in public buildings, or in homes. Since these structural features represent large vertical forms in space, they often become the dominant design feature in an interior space (Figure 14). Stairs in hotel lobbies, for example, are usually in very prominent locations. The actual stair design, however, is surprisingly restrictive and set. The height of riser and its relation to the tread is fixed, and variations for normal vertical circulation are extremely

limited. Matters of detail involve such considerations as whether the stair is open or enclosed, whether it is a bold sculptural form or an airy dynamic shape (resulting from the use of open treads without risers), whether the stair honestly expresses its material (be it wood, steel, or marble), or is wrapped in carpeting. The many detailing possibilities present a real challenge to designers and, unlike mass-produced windows, light switches, or plumbing fixtures, give designers a chance to design in a completely personal or creative way.

Maynard L. Parker



Figure 14: A ramp functioning as the focal element of an interior: the former V.C. Morris Shop, San Francisco, designed by Frank Lloyd Wright, 1948.

Placement of utility fixtures

Components such as heating units, electric outlets and switches, and telephone connections offer no design choice other than the limited selection among mass-produced products and the best placement within the space. The pattern created by the placement of fixtures is as important with walls or any other surfaces as it is for ceilings. A given wall may have doors, windows, electric outlets, switches, air-conditioning registers, and heating units (radiators or convectors). It is the designer's job to deal with all of these components by design, by organization, by placement or elimination, and by detailing. Often, the more bulky components, such as radiators, can be "eliminated" by building the unit into the wall or, in existing, poorly detailed buildings, by creating a "built-in" appearance through the inclusion of some design feature. Radiators or convectors are often housed in neatly detailed enclosures that may run the whole length of a window wall and may at the same time provide an additional surface under the windowsill. Depending on the location, a continuous enclosure may contain some shelving or storage elements, thus making use of the extra space not needed for the actual heating unit (or air-conditioning unit).

In large, nonresidential interiors, the mechanical components are often massive. For instance, the telephone installation needed in an office for several hundred people requires a very large space and a complex installation of conduits and other elements that affect the interior design. The air-conditioning or heating unit for a simple store may be fairly bulky, and again the designer deals with the allocation of space as well as with the mechanical function of the equipment. All of the mechanical equipment for buildings is specified or engineered by specialists, but it is essential that an interior designer have the basic knowledge and understanding to be able to coordinate the various specialties. The many pipes, stacks, and vents that go into a plumbing system, although not exposed and shown as a rule, are of real concern to the designer. Whether architectural components are expressed and detailed, whether they are concealed or built-in, they are incorporated in the design.

Furniture and accessories. To the layman, furniture is the most important aspect of interior design. It is a significant component of design to the professional as well, since it is the most personal and intimate product relating man to a building. It is also personal because it can be moved from one home to the next and handed on from generation to generation, and often furniture takes on important sentimental value. Accessories are even more personal, but they are less significant to the overall effect of the interior, since they are by nature smaller than furniture. Almost anything that people own or collect could be called an "accessory," including functional objects, such as ashtrays, and decorative objects, such as porcelain, glass, or ceramics.

Personal nature of furniture

Although some quite sophisticated furniture existed in ancient Egypt, the use of furniture was rare during the Middle Ages and only became significant in the West during the Renaissance. During most subsequent periods there have usually been close interrelations between architectural and furniture styles and modes of interior design. (That aspect of furniture will be discussed below under *Historical and stylistic developments of interior design.*) The 20th-century pioneers of design and architecture—such as Mies van der Rohe, Le Corbusier, and Marcel Breuer—were not able to find any suitable contemporary furniture available in the 1920s and 1930s when they built structures without historical references. They designed much of their own furniture, and some of these modern "classics" are still very much in demand. Well-designed modern furniture developed in Scandinavian countries in the 20th century out of the long tradition of craftsmanship and design prevalent in those countries. The real beginning of modern furniture design in the United States came only after World War II, and much of it was first developed for nonresidential uses. Charles Eames, George Nelson, and Florence Knoll are among the distinguished American designers who have pioneered furniture design and manufacturing processes. Their furniture primarily was introduced to the public through use in public or work spaces. A large segment of furniture manufacturers, however, has still not been touched by design of any kind, and furniture under such invented names as "Mediterranean" or "Italian Provincial" (both nonexistent historic styles) is still being foisted upon the public.

Use of built-in furniture

Whatever material or manufacturing process may be used, the important criteria that must be applied in furniture are function, comfort, and durability, together with aesthetic considerations. Architects and interior designers often prefer to build in furniture wherever possible, and, indeed, some of the best historic and contemporary interiors contain little movable furniture. An interior without any furniture or accessories would probably appear stark and uninviting, and it is clear that the personal touches possible through selection of appropriate furniture and accessories are very important.

One can use a vast array of decorative objects or plants as accessories. In a way, every accessory used in a home, office, or public space is in some way a part of the total composition, and must therefore be selected with care. No rules exist on what is "proper" other than the basic principles of design that were discussed earlier.

Lighting. Light is one of the key elements of interior design. Most interior spaces constructed in the 20th century are used as much with artificial light as with daylight; because of this lighting has become a very significant tool for the interior designer. There are three major aspects to lighting: function, aesthetics, and health. The latter factor is often ignored, but insufficient illumination can cause eyestrain and physical discomfort. Illuminating engineers have established recommended standards of illumination for various tasks and have also provided rules and standards relating to brightness of the source of lighting and controls for shielding the eye from direct glare. Light can be diffused and can, in general, be controlled very accurately.

Two basic types of lighting are used in modern interiors:



Figure 15: Fluorescent, incandescent, and neon light used to create a particular atmosphere or mood: Ocean Tank, New England Aquarium, Boston, architects and designers, Cambridge Seven Associates, Inc., 1967. The dim light of the aquarium is immediately evocative of the dark, mysterious underwater world.
By courtesy of the Cambridge Seven Associates, Inc.; photograph, Norman McGrath

Types of lighting

incandescent and fluorescent. The former is somewhat redder than daylight but contains all colours of the spectrum. Since fluorescent light has an uneven spectrum, colours tend to appear distorted. A mixture of the two is often the best way to achieve colour accuracy. Some of today's fluorescent lamps are close to daylight accuracy, and manufacturers continue to improve the quality of available lamps. Both types of light can be used in "direct" or "indirect" lighting in interiors or in a combination of these methods known as semidirect or semi-indirect (Figure 15).

Designers and architects strive to build in lighting as much as possible. Recessed lighting, lighting coves, and architectural lighting in general can be controlled much more efficiently than portable lamps.

A good lighting scheme must provide some variety in highlights, shadows, and accent lights to avoid monotony. An even, overall lighting system, such as a luminous ceiling, can be highly efficient, but it lacks character and interest. Most interiors require a certain flexibility for different functions within the space at different times of day and night. In certain interiors, such as stores and shops, lighting becomes a display and sales tool, and in festive spaces, such as ballrooms or theatres, the quality of light can provide sparkle and mood more effectively than any other component of design. One can think of the potential of lighting in terms of the theatre. Some productions are staged without formal sets, yet the changing mood and setting can be suggested by controlled illumination.

Design in portable lamps

Most intimate interiors depend to some extent on portable or fixed (ceiling and wall-mounted) lamps. The design of lamps, especially table lamps for homes, has somehow brought forth a vast array of bad designs, together with a smaller number of good ones. Many lampshades are similarly banal in design, but a shade as such is an excellent diffusor of light and shield against glare. Some lamps and shades are designed for specific tasks, others for accent lighting.

Fabrics. There are three basic aspects that determine appearance and suitability of fabrics for interior use: fibre content, weave, and pattern. Fibres are either natural or man-made. The important natural fibres are cotton, wool, linen, and silk. Although silk has long been considered the most elegant and desirable of all natural fibres, it does not stand up well under direct sunlight and heat and, in general, requires more care than most other fibres. Wool, like silk, is an animal fibre; depending upon its weave, it can be made into extremely strong and beautiful fabrics and is therefore very much in demand for contemporary interiors. Both cotton and linen are made from vegetable fibres and are both durable and pliable. Unless cotton and linen are interwoven with other fibres, however, they are not generally as strong as wools or man-made fibres and tend to be restricted to light-duty interior purposes.

Families of man-made fibres

Man-made (synthetic) fibres in the 20th century abound under a variety of trade names, and new synthetics are continuously being developed. Some of the major families of synthetic fibres are glass fibres, acetate, acrylic and modacrylic, nylon, olefin, polyester, rayon, and saran. The chemical composition and processes used in the manufacture of man-made fibres make possible a variety of specific qualities. Some offer strength and elasticity; some offer resistance to fire, stain, mildew, sun, or abrasion; and some offer resistance to moisture and organic agents, others to crushing and wrinkling.

Many fabrics are woven in a combination of two or more fibres in an attempt to improve the appearance or utility or both. Another factor in selecting or specifying fabrics is the touch of the fabric, or the "hand." Certain fabrics made from man-made fibres seem unpleasant to the touch compared to silk or wool fabrics.

Weaving is an ancient art, and fundamentally there is little difference between the very early handlooms and the power looms found in major textile plants today. The three most common weaves in use are plain weaves, which include basket weaves; floating weaves, which include twill and satin weaves; and pile weaves, which include both cut and uncut weaves. Weaving techniques of lesser importance to interior design include knitting, twisting, forming, and felting.

The pattern of textiles, especially in contemporary terms, is frequently the natural pattern created by the weave of the fabric, although patterns are also created by printing. In traditional textile terms, reference to pattern usually meant a historic style. The history of textiles ranges from early Egyptian and Oriental patterns to the present. Each era has developed fashionable and popular patterns. Contemporary textile designs, for instance, are usually abstract or geometric, but floral and large flowing patterns were also popular in the 20th century.

Colour in fabrics

Colour is one of the most important aspects of fabrics in interior design, inasmuch as the colours of fabrics are frequently the most important areas of colour in interiors. Dye colours can be added to unspun fibres, spun yarns,

or woven textiles. Colour fastness is a major concern to interior designers, for faded fabrics can be quite detrimental to an interior.

Natural elements. No man-made object can equal the beauty found in nature, and it is not surprising that the introduction of natural elements into interiors has always been considered desirable. In spite of their beauty, one cannot arbitrarily introduce a plant, a tree, or rocks, or water into an interior. The foremost considerations must be the location of the space, its climate, and its relationship to the outdoors.

Climatic considerations determine the kind of plant, flower, or tree that can prosper in an interior. The most beautiful plant will not survive long under adverse conditions, and a dying tree or plant certainly offers no decorative advantage.

The location and orientation of interior to exterior spaces is another important consideration in the introduction of natural elements. In warmer climates, it is possible to have a gradual transition between interior and exterior, and plants providing this natural transition will look well and will prosper. In colder climates a real barrier of glass or a solid wall separates the indoors from outdoors, and at best the transition can be made visually.

There are a number of simple devices that make it possible to keep delicate plants and flowers alive under controlled conditions. Greenhouses in all sizes, ranging from window size to room size can be the most delightful areas of an interior, but obviously special conditions and maintenance must be provided. The scale of plants or small trees must be considered. One large indoor tree can be a beautiful accent in even a small space. Too many trees or plants in a small space would be overpowering, unless indeed the space is designed primarily as a greenhouse space or plant room.

Natural elements other than plants and flowers that can be used in interiors are water, rocks, stones, or pebbles, and planting areas in natural soil. For large spaces, usually public buildings, pools or contained areas of water can be extremely beautiful and exciting. Some interior features have been created with running water and small recirculated waterfalls. Sometimes a small area of pebbles with a few plants or carefully chosen rocks can add a touch of real beauty to an interior. Even collections of rocks, minerals, seashells, and other natural elements provide the touch of nature that can make an interior come alive.

DESIGN PROCEDURE

Professional interior-design assignments may range from the design of a small apartment to extremely large and complex jobs such as the planning and design of all of the floors in an office building or the design of all the spaces in a hotel or resort. The procedures vary somewhat from one job to the next and depend upon the size of the design organization, but the following basic outline covers the usual procedures followed by professional designers.

Interview with the client

Preliminary phases. The first step is the interview with the client. This is often a series of conversations and must eventually lead to a mutual agreement. Clients usually have a good idea of their needs and preferences, yet an experienced designer frequently sees some needs not envisioned by the client, and often he must reeducate the client's attitude about preferences. Obviously, the interview must also convince the client that the designer is the right one for his needs. Most established professionals do not commence any design work nor engage in prolonged meetings and conversations without a retainer for their services. Depending upon the scope and complexity of the job, agreements between clients and interior designers range from simple letters written by the designers to lengthy legal documents, covering precisely the services to be rendered, as well as the procedures and responsibilities. The designer makes a survey, including an analysis of the client's present program, and he often prepares a new program. Frequently, for instance, a designer upon surveying existing facilities finds that the redesign of these facilities would be more suitable to the client's needs and more economical than the leasing of a new space or the adding of additional space. More often the situation is reversed: the client does not realize that investing in a major renovation of his space does not permit room for future change or expansion, and upon the design firm's advice new premises are obtained or built. Sometimes there is a question of whether a particular interior of some value or meaning should be restored or reconstructed, and again the experience of the interior designer is needed for those decisions.

When the job involves redesigning existing spaces, at a very early stage the interior designer will require very accurate plans of existing conditions. In many older buildings, there are no up-to-date plans, and the design firm must take exact field measurements in order to obtain plans and elevations for the existing spaces. These plans must also reveal whether walls are bearing (supporting) or whether they can be demolished. The electrical and mechanical system must be carefully evaluated, sometimes by engineers.

The need for up-to-date plans

For large jobs pre-architectural planning and programming can consume many months or even years. Major corporations contemplating major building projects need precise programs, analyses of existing facilities and equipment, and a number of alternate schemes and proposals. Based upon the functions performed by the various departments of a corporation and the interrelation of these departments to each other, designers actually prepare a schematic building shape (such as a high-rise building or a series of smaller structures), including a basic system for offices or other functions.

The final program outline is eventually presented to the client for approval prior to any actual design work. The budget obviously is a paramount consideration. Together with the program analysis, designers must frequently prepare an approximate budget or attempt to make their proposals based upon a budget set by the client.

Among the additional factors that must be considered are availability of materials and furnishings, maintenance of the interior, and the character or appropriateness of the planned scheme. Business interiors often represent large investments for the clients, and a delay of several weeks in the completion of a job, due to the non-availability of products or furnishings, could represent a sizable loss. In public interiors, such as hotels, stores, or educational institutions, the maintenance factors must be carefully analyzed. On a smaller scale, residential interiors must be considered with similar care. Maintenance factors for the floors of kitchens or children's rooms are important.

Maintenance factors

Design and presentation. After the completion of a program and the acceptance of the program by the clients, the actual design work can begin. Designers usually work on many alternative schemes. A single space such as a restaurant or a carefully designed store takes many days of preliminary design studies. As the size of the job increases, the interrelation of individual spaces increases the complexity of these studies, and it is quite likely that the designer will need a rough study model in order to visualize the spaces three dimensionally. Drawing and drafting at that stage is the designer's way of visualizing his own ideas and at the same time putting them in such a form that they can be communicated to his associates for discussion and eventually communicated to his clients. All the aesthetic components come into play at that stage of design, including colours, lighting, and textures, although at the early design stages no precise selection of materials or objects is made. Obviously, this creative phase of interior design is based on thorough research and critical analysis and is not simply the result of a sudden flash of inspiration.

Once the designer or the team of designers feels that a scheme has been arrived at within the stated objectives, a preliminary presentation will be prepared. Although a competent designer will try a number of possible schemes for every job, he will, as a rule, decide which of the many ideas he explored in rough form is the most successful and that will be prepared for a preliminary presentation. For important commissions, such a presentation might consist of a number of sheets or presentation boards showing plans, elevations, sketches, and renderings, and, in many cases, models as well. Most clients are not

The preliminary presentation

trained to visualize space from plans and elevations, and perspective sketches and renderings are necessary to fully explain a scheme. At the preliminary presentation the specific colours, furnishings, and details are not resolved yet, since the aim at that stage is to obtain the basic approval from the client.

Final drawings and specifications. If a preliminary presentation has been completely accepted, the designers can proceed to the final design stages. If changes have to be made, another meeting (or meetings) with changed presentations may be necessary.

Preparation of final drawings

The next stages of the design may consist of a series of drawings done by professional draftsmen or by the interior designer himself, if he works as an individual. Depending on the type of job, final drawings may consist of just a few sheets or a very large number of drawings. Plans, elevations, details, sections, and specifications are the language of architectural and design offices, and they are prepared with carefully drawn dimensions and notes for the many contractors who carry out the actual construction. Certain drawings may be done by subcontractors or related trades; for instance, the air-conditioning system is usually designed by air-conditioning engineers, and the duct work must be designed in connection with the lighting system in order to assure that lighting fixtures do not conflict with ducts. Similarly, mechanical equipment—such as heating or plumbing pipes, telephone cables, and electrical lines—must be coordinated to avoid conflicts and problems. Before outside firms or subcontractors become involved, the designer or design firm usually prepares the design drawings with sufficient information to enable various contractors to submit bids. Almost all major jobs are sent out for bid to several contractors, in order to provide the client or the designer as his agent with a series of competitive estimates.

On complex and costly design commissions a final and elaborate presentation may be prepared after the acceptance of the preliminary presentation. This might include very carefully drawn perspective renderings in colour. Many presentations include scale models and may consist of nothing but carefully crafted models.

Selection and specification of furnishings

Together with the preparation of final drawings, interior designers begin the process of final selection and specification of all furnishings. The process of selecting and ordering fabrics, furniture, lighting, and all other furnishings requires a thorough knowledge of available products. In large cities there are often hundreds of sources, but, in spite of the vast product choice available, it is not always possible to find just the right fabric or just the right piece of furniture. In such cases interior designers may have to design special furniture, floor coverings, lighting fixtures, or fabrics. Most interior designers are familiar with quality products and maintain within their offices samples and catalogs of furnishings that they consider of merit. The products that have been selected by the designer are usually submitted to the client either as part of the original design presentation or in a separate approval step. Methods of placing purchase orders vary. Many design firms and individual designers prefer to limit their activity to selection and specification and arrange to have the client's purchasing office place the orders. In other cases the designer places the purchase orders but then submits the invoices to the client for direct payment. In either case ordering and specifying is an exacting task. Delivery and availability is an important concern that the interior designer is responsible for. If a hotel is scheduled for opening at a specified date, it may be necessary to place orders for furniture and furnishings as early as two years before completion date in order to assure delivery on time.

Building supervision

Construction. The actual building of the interior, be it a renovation or a new construction, needs considerable supervision by the designer, although constant on-site supervision is not always required. For an office or residence, a few visits may be sufficient. The thoroughness of working drawings and details influences the degree of supervision that is needed: the more complete the drawings and specifications for a particular job, the less time must be spent on the site during the building stage.

In spite of the fact that the workers are usually highly skilled craftsmen, there are questions that can only be answered on the site, and there are always unforeseen problems that require changes or on-the-spot decisions. Many interior designers have considerable understanding of construction and building technology, can communicate with tradesmen intelligently, and are able to offer valuable advice and suggestions. The situation can also be reversed. Many construction workers are very skilled and knowledgeable and are able to offer suggestions that designers are happy to accept. The supervision must proceed through all stages of a job. Knowledgeable designers spare no effort to see that every phase of the job is done in the best possible way.

As with other furnishings, interior designers select, commission, or purchase artwork, plants, and accessories. In residential interior design, clients usually own many of these things or will certainly be involved in the selection and purchasing, but in interior design for commercial or public spaces this responsibility is in the hands of the designer.

From the foregoing discussion, it will be clear that the design of large interior jobs involves many detailed considerations from the inception to the completion. For this reason most large design firms dealing with hotels, governmental or institutional clients, or large business firms have developed work sheets and checklists for all aspects of the work. Each phase of a job is usually under the supervision of a job captain or chief designer, and each checklist or form is controlled and checked repeatedly in order to assure that everything has been considered and that the job is moving smoothly to completion.

Work-sheets and checklists

KINDS OF INTERIORS

Although the foregoing sections have mentioned different kinds of interiors, in reference to both aesthetic and physical components of design, there has been no specific discussion of different design considerations for varying interiors. The aesthetic criteria suggested in earlier sections are subject to considerable variation, depending on the kind of interior involved.

Residential interiors. Residential interiors are obviously much freer and much more personal for both the interior designer and the occupants than other types of interiors. In fact, homes that have been designed unconsciously by creative occupants without any standard decorative rules are often the most beautiful ones. Certain planning and functional considerations are constant in any residence, and, although these too may be ignored by the occupant who wishes to be strongly individualistic, they can provide at least basic guidelines.

Relation of one space to another

The planning of modern houses or apartments must take into consideration the location of certain needs in relation to others. The dining space should be near the food-preparation area, and the food-preparation area should be accessible to the entrance used to bring in food supplies and remove waste. Access to children's sleeping areas should not be through the adults' living spaces. Access to bathrooms should be close to the bedroom areas and should not be through living or dining spaces.

The furniture arrangement for a living space must take into account the occupant's life-style and preferences. If a space is planned for young people, no seating might be provided other than the floor, but, for the more conservative or older occupants, comfortable seating for conversation and other activities is essential. Open-plan houses (living, dining, eating facilities without separate rooms) work splendidly and beautifully for some people but might not be the ideal answer for a family with many children and a desire for privacy at the same time. The special storage needs that must be considered for many homes vary from bookshelves to storage areas for bicycles, from facilities for recorded music to storage of sporting equipment. Such facilities can often be added by interior designers, if not provided by the architect.

There are several types of residence, and each one may require a different approach, partially based on economic considerations. The private house owned by the occupant warrants not only built-in designs and other permanent design features (lighting, flooring, etc.) but, in general,

lends itself naturally to anything within the imagination of the designer and the budget of the owner. Cooperative apartments are prevalent in larger cities, and those that are bought outright by the owners can be designed and changed as long as the structure of the building is not tampered with. A different approach is usually called for in rented apartments or houses. Major changes and special furniture and other built-in features would be considered a poor investment by the client and would, as a rule, be frowned upon by the landlords.

Residences as status symbols

In the past, professional help for residences has been basically reserved for wealthy clients. The residences involved were often status symbols, and the furnishings were to a large extent traditional furnishings and antiques. The best of such ornately designed homes are authentic, museum-like interiors, which indeed only the very affluent can afford. (Most status-conscious interiors, however, consist of reproductions and imitations and have little to do with good design.)

Today, instead of being limited to the service of the wealthy, the designer has a widening and important opportunity in a totally different aspect of residential interiors: mass housing and low-income housing. Although only in recent years have some designers involved themselves in this area, with an increasing concern on the part of both government and private enterprise for the effect of environment, the field should offer a growing opportunity for challenging creative work. Such designers, as well as helping to create more liveable spaces for those with limited housing budgets, can also be of great help in assisting occupants to choose simple, sturdy, attractive, and functional furnishings. A major problem for many people, on a variety of income levels, is the high cost of furnishings; mistakes in judgment are too costly to be discarded and thus must be endured. The help of professionals can minimize this problem and also protect low-income families from being induced to buy installment-plan furnishings of poor quality and design.

Public interiors. *Space planning.* Although many designers are engaged in residential interior design, there has been a marked shift away from that field since 1950, and more designers than ever work in the design of public, institutional, and commercial spaces. Space planning for business firms, governmental agencies, and institutions is a significant aspect of office design and is concerned primarily with planning, allocation of spaces, and interrelations between offices, departments, and individuals. The aesthetic or design phase varies with the degree of importance attached to offices by the clients. In a large firm, the clerical, accounting, or filing areas tend to be well designed in terms of lighting, efficiency, space, and function but have few frills or design features. The executive offices, reception areas, and conference rooms, on the other hand, are frequently elaborately and luxuriously designed, since they serve as images for the corporations as well as status symbols for their occupants. Decisions relating to size of offices and their furnishings are basically arrived at through functional considerations. An executive frequently must seat groups of people in his office. A department manager or clerk will rarely need more than one or two extra chairs.

Pre-architectural planning has taken on such importance that many design firms provide this service. Through careful study and analysis, standards of typical offices, relationships of offices and departments to each other, the need for flexibility and storage, and many other aspects of work within a given business can be arrived at, and such a study then becomes the program for the actual design of a new building or premises. When truly large firms or governmental agencies are involved, space studies preceding the actual design may take several months or even years.

Office landscape

A rather recent innovation in office design is known as office landscape (from the German word *Bürolandschaft*). Above, in *Modes of composition,* it was noted that the appearance of a "landscaped" space might seem chaotic. Actually, however, the system was developed in the 1960s by a German team of planning and management consultants who made intelligent use of computer technology to arrive at predictable relationships between persons and departments in a given organizational structure. Office landscape also takes into consideration the high cost of building and the continuous need for change in large corporations. The solution offered by these planners was not to build the traditional permanent walls and private offices but to arrange a large open space in a purely functional plan. Divisions between people and departments are created by free-standing screens, and plants are often used to divide and enhance space. Office landscape has been used in several major installations in the United States, following considerable popularity in Europe, but there are skeptics who question the basic claims of office-landscape supporters that less space is required and that the resulting democratization creates a better spirit and working relationship among staff members.

It is interesting to note that even in conventional office planning there is controversy about whether or not the occupant of an office should be involved in its design. Designers tend to insist on making all decisions, and management usually supports that point of view, yet psychologists, among others, counsel that a greater involvement of the individual with his own personal environment would be desirable.

Governmental interiors. A notable characteristic of interior design for public buildings—such as court rooms, assembly halls (on all levels of government including the United Nations), city halls, and cultural buildings—is that the consumer is excluded from participation in decision making. Another is that in all cases the interiors try to present a very definite image or symbol. Governmental buildings, especially in the past, were designed to present a solemn, awe-inspiring, majestic, and even slightly ominous look, both in their architectural composition and their interior treatment of spaces. For centuries, marble, stone, lofty ceilings, and imposing architectural elements have been traditional.

The image of government buildings

Institutional interiors. Schools, hospitals, and universities are examples of institutions now extensively using the services of interior designers and architects. Many universities have staff designers dealing with the institution's many design needs, from office spaces to dormitories. Certain institutional needs, such as operating rooms in hospitals, are strictly functional, yet the patients' rooms and many other hospital facilities are very much within the scope of interior design. Until recently, however, such involvement was not prevalent, and it has been common to refer to a sterile, dull-looking space as "looking like a hospital." A greater recognition of the influence of the environment upon human behaviour has brought about increased emphasis on interior design for all kinds of institutional interiors. Indeed, even though up to now little work has been done by designers in penal institutions, it is a safe prediction that in a short time there will be considerable concern for the environmental qualities of these institutions, as well.

Commercial interiors. Contemporary designers are much involved with commercial spaces—such as stores, hotels, motels, and restaurants. Many designers and design firms specialize in highly specific spaces such as restaurants, and others may become specialists in the design of showrooms for the garment industry. Frequently, the design of a restaurant, shop, or hotel must be keyed to a theme. It might be a nautical theme for a yacht club or a theme based on the artifacts of the particular region in which a hotel is located. Obviously, all commercial spaces must be designed in a highly functional way. A store with a beautifully designed interior will fail if it does not work for circulation of customers, for display, for storage, and above all for sales. Some of these functional needs create difficult design problems. A hotel or motel room, for instance, must be designed for use by individuals, couples, and family groups. Maintenance is also an important factor in the design of commercial spaces.

Use of a theme

Religious interiors. Religious architecture is heavily influenced by symbolic concepts as well as by the ritual and traditions of a particular faith. Designers of religious interiors must, therefore, base their approach on a set of rules preceding all other design considerations. The simple and modest Quaker prayerhouses, for instance, express the

tenets of that faith as clearly as some of the richly appointed Roman Catholic and Eastern Orthodox churches.

Industrial interiors. Industrial interiors do not usually involve interior designers. There are, of course, many industrial spaces, such as workshops, laboratories, and factories, that have been planned by architects and designers, and there are a few that have stressed some aesthetic considerations. By and large, however, industrial interiors are created as strictly functional spaces. For this very reason, some of these spaces are quite beautiful. This may sound paradoxical, but, like the modern bridge or airplane, they can be extremely handsome without the conscious attempt to create beauty.

Special interiors. Although an attempt was made to classify the kinds of interiors that are the prevalent concern of interior design, there are many kinds of special interiors that at times fall within the larger field of environmental design and that do not fit into a particular category or even a professional subspecialty. Transportation design may be part engineering, part industrial design, part architecture, and part interior design. Interiors of ships are certainly interior design, but the interiors of automobiles, aircraft, and trains are often a combination of many specialties. The advent of large commercial aircraft has taken the aircraft interior out of the area of the strictly functional, and, indeed, the introduction of these large planes has seen an intense competition among the airlines to create spaces that go beyond the concept of mere seating. Also included in transportation design are the terminal buildings associated with air, road, and water transportation systems.

Transportation design

A less spectacular example is the field of exhibition design, another area of design having interfaces with other fields, including, in this case, graphics and advertising. Related to this field are museum design and exhibition and the preservation and restoration of historic buildings.

It is clear that any man-made interior or exterior space is influenced by design or its absence. More important than a listing of the various kinds of special interiors is the underlying fact that designers are becoming involved in all aspects of the environment. (A.A.F.)

Historical developments

The art of interior design encompasses all of the fixed and movable ornamental objects that form an integral part of the inside of any human habitation. It is essential to remember that much of what today is classified as art and exhibited in galleries and museums was originally used to furnish interiors. Paintings were usually ordered by size and frequently by subject from a painter who often practiced other forms of art, including furniture design and decoration. Sculptors in stone or bronze were often goldsmiths who did a variety of ornamental metalwork. The more important artists had studios with assistants and apprentices and often signed cooperative work. Many architects also designed interiors, including the accessories—furniture, pottery, porcelain, silver, rugs, and tapestries. Paintings often took the form of cabinet pictures, framed to be hung on a wall in a particular position, such as over a door. Murals were painted on a diversity of subjects; during the period of the Baroque style in the 17th century, murals sometimes were painted to look like an extension of the interior itself, making it appear more spacious. Mirrors were employed for the same purpose of adding space to an interior.

The deliberate use of antiques as decoration was unusual in most periods. Generally, in older houses elements of the previous decorative scheme were relegated to less important rooms when new decoration was undertaken to bring an old interior into line with current fashion. In this way many antiques have been preserved. The art market has existed from the earliest times for the purpose of providing both new and antique works for the decoration of interiors, but in early times the market in old work was usually limited to paintings by admired masters and goldsmith's work.

Museum collections of furnishings

Only within the recent historic past have any interiors but those belonging to the rich and powerful been considered worthy of consideration. Still more recent is the collection of the interior furnishings of the past by museums and galleries, where they are studied in scholarly isolation. The segregation of such objects in galleries, however, has led to an increasing misunderstanding of their original purpose; and the division of the arts by museum curators into the fine arts and the decorative (or industrial) arts has helped to obscure the original functions of interior furnishings.

To some extent the present attitude has resulted from the rise of the specialist collector since the 1840s. Porcelain and silver, for instance, no longer fulfill their original purpose as part of the household furnishings but are collected into cabinets, since they are so precious. Similarly, the small porcelain figures of Meissen, which were originally part of a table decoration and an integral part of a service, are now too highly valued to be so used.

ORIGINS OF INTERIOR DESIGN

The notion of interior design historically has arisen as part of a settled agricultural way of life. The tents of nomadic peoples were hardly suitable for the more permanent forms of decoration. Among Central Asian nomads, however, carpets and rugs have been employed to decorate and provide comfort in tents and portable dwellings, usually taking the form of coverings for floor and bed, and these have been the principal form of art of the peoples concerned. The oldest nomadic carpet, found in Central Mongolia, dates to the 5th century BC, but geometrically patterned stone reliefs from Assyria in the 7th century BC are thought to be based on earlier carpet patterns.

Hunting peoples living in caves decorated the walls with paintings as early as 20,000 years ago, but these were almost certainly votive paintings rather than decoration, and no trace of movable furniture has survived.

Primitive peoples. Although the practices of present-day primitive peoples sometimes shed light on the historical origins of those practices, there is too little art and decoration in such communities today to illuminate the beginnings of interior decoration. No clear-cut progressions of styles, like those that occurred in Europe, can be identified except among peoples who could hardly be regarded as primitive, such as the former civilizations of South America or the Benin culture of Africa. Nevertheless, even the poorest and most primitive peoples devote some time to the production of works that give them pleasure, and these works often are employed to decorate interiors. Primitive painting often consists of a series of abstract patterns, such as that on the pottery of the Pueblo Indians. Furniture, such as wooden stools, usually has some ornamental carving. Basketwork, wooden vessels, and pottery are decorated with abstract geometrical patterns, and an insistence on symmetry is the rule. Since most of these patterns—especially those to be found in basketry and textiles—bear no resemblance to natural forms, they probably arose from the nature of the techniques employed in making the objects in question.

Connotations of ornament

Ornament based on natural objects more or less realistically depicted probably had a magical connotation; animals, for instance, are intended to promote success in hunting. Even the most abstract and geometric of motifs have a symbolic meaning, which can be interpreted by those who know the key, and this meaning is almost always magical. There are few objects or motifs that do not have some meaning, and the making of objects that have no other purpose than the pleasure taken by their creator in executing them is very rare.

Origins in Western antiquity. Excavations in ancient Mesopotamia and Egypt suggest that the earliest equivalent of furniture consisted of platforms of bricks, which served as chairs, tables, and beds, no doubt spread with textiles or animal skins. There is also good reason to think that walls were painted and, in the case of more important buildings, decorated with mural paintings. Movable furniture first occurred only in the most important residences, such as palaces, and in public buildings. Furniture is of considerable antiquity, though it is known, for the most part, only from wall paintings, sculpture, and vase paintings. Some furniture survives from ancient Egyptian tombs from about 3000 BC in the form of beds, chairs, tables, and storage chests. It is in such furniture that dec-

oration is first seen—in the leg of the bull and the lion employed as a furniture support, especially for beds. It is from this point in the ancient past that the development of interior design can be traced historically.

INTERIOR DESIGN IN THE WEST

Ancient world. *Egypt.* In contrast with the monumental tombs and temples of stone, many of which remained intact to the 20th century, Egyptian houses were built of perishable materials, and, therefore, few remains have survived. Sun-dried or kiln-burnt mud bricks were used for the walls; floors consisted of beaten earth, and a thin coat of smooth mud plaster was often used as an internal wall finish.

In its simplest form the applied decoration was a plain white or coloured wash, but, in larger houses, patterns in varying degrees of elaboration were painted on the plaster. Rush matting was hung across most internal door openings and used as screening inside the small, high windows. It is probable that decorative wall hangings and floor coverings were made of rushes or palmetto woven into a pattern, since painted representations of such hangings have survived from 5th-dynasty tombs at Saqqārah. In the workmen's village of Kahun, built in the 12th dynasty (*c.* 1900 BC), some of the more well-to-do houses contained rooms decorated with brown-painted skirting, one foot (0.3 metre) high, then a four-foot (1.2-metre) dado (the lower portion of wall that is decorated differently from that above it) striped vertically in red, black, and white. Above this the walls were buff coloured with brightly painted decorative panels in the more important rooms, and ceilings were also often of painted wood. It may be assumed that the lavish tomb decoration of all periods was basically derived from the domestic interiors of their time.

Use of natural forms for decorative motifs

Many Egyptian decorative motifs are stylized from natural forms associated with the life-giving Nile. The lotus bud and flower, the papyrus, and the palm appear constantly with borders of checkered patterns or coiled, ropelike spirals, giving an air of space and elegance. The palace of the pharaoh Akhenaton and other large houses at Tell el-Amarna (*c.* 1365 BC) reflect a tendency toward naturalism in their ornamentations. Akhenaton, his queen Nefertiti, and their daughters are frequently represented, usually grouped affectionately together. Other painted panels show animals and birds with twining borders of vegetation. Molded, coloured, glazed ware was introduced to give a brilliant inlay of grapes, poppies, cornflowers, and daisies, all in natural colours. The use of square ceramic tiles as a wall surfacing was uncommon but not unknown. Primary colours were the most common, a brilliant yellow being among the most frequently used, but terra-cotta, gray, black, and white were all added to give contrast. Even floors were delicately painted to represent gardens or pools. One of these at Tell el-Amarna shows a rectangular tank with swimming fish and waterfowl, bordered with lotus and papyrus marshland, with an outer band showing more birds and young cattle in the meadows beyond. Furniture ranged from the simplest benches and ceramic pots to beautifully designed chairs, small tables, and beds in the homes of the rich, where many vases, urns, ceramic, wood, and metal utensils evince a fastidious, luxurious way of life.

Mesopotamia. Very little furniture survives from ancient Mesopotamia, principally because climatic conditions are not conducive to the preservation of wood. What is known has been learned principally from reliefs and cylinder seals. Furniture mounts of bronze and ivory have been excavated, however, and fragments of furniture were uncovered in the royal tombs at the city of Ur, in ancient Sumer. In quality of craftsmanship and decoration, Mesopotamian furniture was comparable to that of Egypt.

The mud-brick houses of the Sumerian and Old Babylonian periods in the Tigris-Euphrates valley resembled their modern counterparts in their rectangular outline and the groupings of rooms about a central court, which was either roofed or open. In most houses, decoration probably was confined to a wide black or dark-coloured skirting painted in diluted pitch with a band of some lighter colour above. Door frames were sometimes painted red, probably as a protection against evil influences, and where doors were used they may have been of palm wood. The poorer houses were simply whitewashed.

Assyrian relief panels and friezes of ceramic tiles

In the most elaborate Assyrian palaces the main decorative features were panels of alabaster and limestone carved in relief, the principal subjects being hunting, ceremonial, and war, as in the palace of the warrior king Sargon II at Khorsabad (705 BC). Panels and friezes of ceramic tiles in vivid colours decorated the walls inside and out, and it is evident that this brilliance of colour was a feature of much Assyrian and Babylonian decoration (Figure 16). Carved stone slabs were used as flooring, with typical Mesopotamian rosette and palmette (stylized palm leaf) borders. Occasionally, Egyptian lotus motifs also appear.

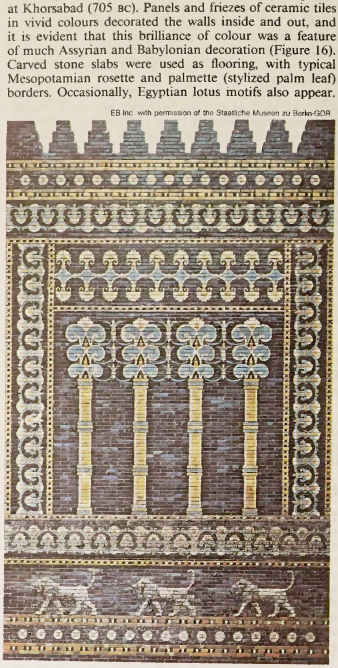


Figure 16: Brilliantly coloured glazed brick decoration, facade of the throne room, palace of Nebuchadrezzar II, Babylon, *c.* 600 BC.

Vigorous and warlike figures characterize both Assyrian and Babylonian work, and the standard of execution was extremely high. Naturalistic detail was often engraved on the surface of the figures and animals, which themselves were in relief. After the Persian conquest (539–331 BC) this vigour declined. The palaces built by the Persian kings Darius and Xerxes I at Persepolis show a lighter use of animal figures. Glazed and enamelled tiles were used on the walls, while timber roof beams and ceilings were painted in vivid colours.

Crete. The most important buildings of the pre-Hellenic Minoan and Mycenaean periods were the citadel complexes, housing the entire court of the ruler. The palace of King Minos at Knossos in Crete (*c.* 1700–1400 BC) gives evidence of a small but sophisticated society with a taste for luxury and entertainment and a corresponding skill in applied decoration. Frescoes (paintings executed with water soluble pigments on wet plaster) and some panels of painted relief decorated the walls of living rooms and ceremonial rooms, which were grouped asymmetrically round a series of courtyards (Figure 17). Many aspects of Cretan life were depicted, the recurring theme being the acrobatic

Cretan use of painted decoration



Figure 17: Frescoed throne room, palace of King Minos at Knossos, Crete, *c.* 1700–1400 BC.
Bernard G. Silberstein—Rapho/Photo Researchers

bullfighting on which a religious cult was probably centred. Even the backgrounds of friezes and panels, which depicted many-coloured painted birds, animals, and flowers, were given an effect of movement, being divided into light and dark areas. Plain dadoes and borders provided an effective foil and gave articulation to the interiors.

As seafarers, the Cretans could import a rich variety of materials for building and decorative purposes; a wealth of ideas can be seen in the fine pottery, carved ivories, and beaten gold, silver, and bronze with which their palaces were ornamented.

The pottery and metalwork of the Minoans was technically in advance of other Mediterranean peoples of the time, and they were especially expert in firing such large pottery objects as storage jars and baths. Some furniture, especially storage chests, was made of terra-cotta. A chalice made of obsidian, a volcanic glass about as hard as jade, could only have been shaped by grinding with an abrasive such as emery procured from Cape Emeri on the island of Náxos; the form was apparently based on metalwork. Excavations have proved the existence of an advanced sanitary system, with baths either of marble or terra-cotta.

Greece. A period of so-called dark ages in Greece followed the destruction of Knossos in *c.* 1400 BC, but Cretan civilization had already influenced the mainland before then. Small terra-cotta models of furniture and fragments of tables and chairs dating from as early as 1350 BC have been found. Homer's epic *Odyssey*, dating from the 9th–8th century BC, speaks of a chair inlaid with ivory and silver, and sheet copper was used to sheathe beams and architraves. The description of a bed reveals it to have been a rectangular wooden frame with coloured leather thonging, like the usual Egyptian bed, and inlaid with silver and ivory. At this time also, wooden vessels were decorated with sheet-gold ornament with repoussé work (ornament in relief made by hammering the reverse side).

Little or no Greek furniture survives from the classical period (5th century BC), but there is ample evidence that it was well constructed and elaborately decorated. The large number of surviving painted vases are a valuable source of information about many aspects of Greek life, and furniture of all kinds—chairs, tables, day couches used for dining, and a large number of accessories—can be identified. These paintings, in fact, were among the major influences on the French Empire style of the early years of the 19th century. Egyptian influence can be traced in some of the early pieces of furniture, an example being a type of chair having a single leg with a lion's head at the top and a single paw at the bottom. This also was to be a favourite theme of the Empire style.

Hellenistic mosaic floors

In the Hellenistic period (323–31 BC), domestic comfort and decoration were considered once more. Mosaic floors were an important decorative device, originally made of pebbles as at Olynthus but later developing into the black-and-white or coloured mosaics that were widely used throughout the Roman Empire (see the section *Mosaic* below). A central, finely designed panel with realistic motifs and a wide, more coarsely executed border of scroll or key patterns acted as a focus for the arrangement of furniture, which was still limited in quantity.

Rome. Much more is known about Roman interior decoration, and Roman furniture was based on earlier Greek models. From the beginning of the Christian era the predominant Western style was that derived from ancient Greece by way of Rome. Classical styles were based on mathematically expressed laws of proportion that were applied not only to buildings as a whole but also to much of the interior decoration.

Roman interior decoration is known both from literary sources, such as Pliny's *Natural History* and the *Histories* of Suetonius, and from excavations, such as those that uncovered the remains of the Golden House of Nero soon after 1500 and those at Pompeii and Herculaneum in Italy in the 18th century.

There are many misconceptions about the decoration of the period, most of which date from the 18th century and the classical revival that began soon after 1750. Many excavated bronze objects, including statues, and any bronze that remained above ground, such as the roofing of the Capitol, were melted during medieval times for new work, since bronze was a scarce and expensive metal. This led to the assumption that marble predominated, which is not necessarily true, especially in the case of statuary. Time and exposure to the weather has removed the colour from much of the marble that has survived, but in classical times it was commonly painted and sometimes gilded. Wall paintings at Pompeii and Herculaneum are ample testimony to this. Wall decoration began there about 150 BC, and, by about 80 BC, plastered walls were being made to look like masonry. Such decoration was combined with the true architectural features—*e.g.*, doors and pilasters (flattened columns attached to the wall). The panels are painted variously in yellow, black, magenta, and red, with some imitation marbling indicating an earlier custom of applying marble veneers. Rich colour was also supplied by superbly executed mosaic floors, elegant couches with coloured cushions, and bronze tripods and lamps, such as in the cubiculum of a villa at Boscoreale near Pompeii preserved in the Metropolitan Museum in New York City (Figure 18).

Roman wall painting depicted columns, niches, and open windows with elaborate imaginary views and figures beyond. Painted ruins, such as those in the Villa of Livia, Rome, were the precursors of the 18th- and 19th-century Romantic taste in western Europe.

It has been said that Augustus, who was emperor from

By courtesy of the Metropolitan Museum of Art, New York, Rogers Fund, 1903



Figure 18: Frescoed room, from a villa at Boscoreale, near Pompeii, 1st century AD. In the Metropolitan Museum of Art, New York City.

Roman marbles and other decorative stone

27 BC to AD 14, found Rome of brick and left it of marble, and certainly the interior decoration of imperial Rome expressed the emergence of the city as a world power toward which flowed much of the wealth of the empire. Exotic marbles began to be imported, and brick walls were faced with polished slabs of white and coloured stone. In the more luxurious interiors or for special purposes, obsidian, a natural volcanic glass dark green or purplish-brown in colour, and copper-green malachite were occasionally to be found in the capital. A limited amount of window glass—mostly small, thick, and discoloured panes—was used, for sheet glass was difficult to manufacture. Large translucent crystals of selenite (a kind of gypsum) were sometimes employed to admit light.

Some of the large houses contained a picture gallery, known as the *pinacoteca*, for the display of easel pictures. These have now virtually disappeared, but mural paintings are fairly common. Pictorial decoration for floors and walls was supplied by mosaics, the picture built up of small fragments (tesserae) of coloured stones, mostly marble, or of small pieces of coloured glass backed by gold foil to increase its reflective power. The subjects are very diverse. Floor-mosaics in dining-rooms were sometimes decorated with simulated fragments of food, as though they had dropped from the table.

Roman furniture was made of stone, wood, or bronze. Villas were largely open to the air, and stone benches and tables were common. Wooden furniture has not survived, but bronze hardware for such furniture is well-known. Buffets with tiers of shelves were used to display silver. Tables were often made of exotic woods and veneers, with ivory, bronze, or silver trim. Tortoiseshell veneers were popular. The dining couches, which replaced chairs, were richly decorated, often with gilded silver or bronze. Chairs followed earlier Greek forms, and while no fixed upholstery was provided, cushions were plentiful.

Roman textiles

The art of tapestry came to Rome from Egypt, where the craft was an ancient one. Few Roman textiles have survived, and those have mostly been found in Egypt and were probably made there. Rugs woven on a linen foundation were imported from Egypt, and fabrics, including rugs, were imported from the Near East. The richest carpets came from Pergamos, in Asia Minor, and were the most highly valued. They were probably woven with gold and silver thread. Nothing survives of these rich textiles because they were all burned long ago to extract the metal. Roman walls were hung with tapestries, and pillars were decorated with textiles. Silk was imported from China until the time of Justinian, in the 6th century, when silkworms were clandestinely brought from East Asia and the industry was established in Europe.

The Romans were highly skilled glassworkers. Domestic glass was made in large quantities, both utilitarian and decorative, and factories were established for the purpose. Mirrors, however, were normally made of polished bronze or silver; if glass mirrors existed at all, they must have been very small.

The amount of bronze employed in household equipment of all kinds was vast. Small pieces of furniture, such as stools, were made wholly of bronze, and a few specimens have survived. Saucepans were made in factories, some bearing what appears to be the trademark of a swan. Lighting fixtures were also made in quantity, of prefabricated parts, and they played a large part in the decoration of the interior. By the 1st century AD enormous quantities of silver went into the making of such objects as large and heavy platters displayed on the buffets. Bowls and similar pieces of hollow ware were commonly decorated with repoussé ornament, less often with engraving, which is usually to be found on the backs of bronze hand mirrors. Antique silver commanded a high price.

Statuary in bronze, from Etruscan sources or looted from Greece and the Greek colonies, decorated the more important interiors. The theatre of Scaurus, for instance, housed 3,000 bronze statues. Some Roman statues have been excavated at Pompeii and elsewhere, but most were remelted. Only one Roman bronze statue has remained above ground in Italy since it was made—the equestrian Marcus Aurelius in Rome.

Pottery was not among the luxuries of ancient Rome. Vessels such as storage jars (amphorae), lamps, bricks, pipes, and architectural ornament were made in factories. Pottery for the table was usually of the so-called Samian ware, although it was made in many other places than Samos; this had a red polished surface and, often, molded relief decoration reminiscent of contemporary silver. Tableware, too, was made in factories and often marked with the name of the potter. Pottery vases of fine quality were made in imitation of those of Greece. They include most of the familiar Greek types, especially the krater (with a large round body, large mouth, and small handles), although the form often varies. The decoration is principally of the red-figure type (black with decorations in red) but is usually much more elaborate than on the Greek originals.

Classical themes and motifs

Themes of decoration are many, and most come from Greek sources. They became part of the vocabulary of classical ornament that was employed during later classical revivals, such as the Renaissance and the Neoclassical movement of the 18th century. The acanthus leaf is by far the most common, and it was in almost continuous use from the 5th century BC in Greece to the 19th century in the West. The Greek and Byzantine acanthus leaf is inclined to be stiff and formal; the Roman and Renaissance form is much more natural. The vine-leaf and grapes motif is also common, and the palmette occurs especially on painted vases. The ivy, laurel, olive, and honeysuckle (anthemion) are usually to be found as frieze ornament, sometimes in stylized form. Festoons, garlands, and swags of laurel were common decorative elements in relief sculpture. "Cable," or "twisted rope," a kind of plaited ornament, was often used for the same purpose. Rosettes—stylized simple roses with equally spaced petals—were widely used. Originally an Assyrian design, they have continued in use to the present. Egglike forms alternating with tongue- or dart-shaped ornaments originally were a carved stone architectural ornament; they were taken over in later times as part of interior plasterwork.

The lion was very popular, especially the mask and paws, and was employed over a long period, as late as the 19th century, as a furniture ornament or as a door-knocker or handle. Mythological animal forms included the griffin and the chimera, both of Mesopotamian origin, and the sphinx, from Egypt and Corinthian Greece. The head of the ram, a sacrificial animal, commonly ornamented altars and candelabra. The ox skull and horns occur during Roman times, but not often thereafter. The eagle, representing Jupiter, was the symbolic motif of the Roman legions. The human mask surrounded by foliage was common and is usually derived from the masks employed in the theatre or from the head of Medusa, which was especially used as a shield ornament. Atlantes and caryatids, male and female human figures, respectively, were originally used instead of plain columns on building exteriors but were later employed for a variety of ornamental purposes—for example, as part of the decoration of some Renaissance cabinets of architectural form. Trophies were always popular. Weapons arranged in a pattern were carried in the Roman triumphs and later sculptured on monuments. This classical form of ornament was later extended to other groups of implements: in the 18th century, for instance, rustic trophies were formed by grouping agricultural implements, such as spades, beehives, and rakes, into a decorative pattern, and musical trophies were made of musical instruments for the same purpose.

A common type of decoration surviving especially in Pompeii is the frieze of small putti, or cupids, in a variety of guises and at work at a large number of different tasks. These persisted in popularity until well into the 18th century, when porcelain figures of putti in disguise or in an allegorical pose became common. They were also painted on furniture or as part of wall decoration.

Grotesques

Equally popular, but remaining virtually unknown till the discovery of the Golden House of Nero *c.* 1500, are the ornamental motifs known as grotesques (because they were found below ground in a "grotto," a word that strictly means an excavated chamber containing murals). Roman grotesques were fantastic figures, human and animal, that terminated in leafage (usually the acanthus leaf) or in a

fishtail, in conjunction with floral and foliate ornament and arabesques. Revived by Raphael about 1517 for the decoration of the loggia of the Vatican, these motifs became widely popular, in many different forms, from the first decade of the 16th century until late in the 18th.

Middle Ages. From the fall of Rome, when the city was finally sacked by Odoacer in 476, to the 15th century, when the Renaissance was already well advanced, information about the decoration of interiors is scarce. Its history has to be pieced together from surviving objects and illuminated manuscripts.

Byzantium. The capital of the Eastern Roman Empire, Constantinople (formerly called Byzantium, later Stamboul, presently Istanbul) was a convenient meeting place for East and West. It felt the influence of Persian art and transmitted it to early medieval European Christian styles. Most surviving Byzantine interiors are ecclesiastical, although secular wall paintings and especially mosaics continued to be popular. The Iconoclasts of the 8th century, however, not only proscribed the making of images but destroyed most of those already existing. Ivory carving was highly developed, and furniture was inlaid with ivory plaques and decorated with carvings. Goldsmith's work, which had existed in large quantities in ancient Rome, was equally popular in Constantinople. Decoration was usually of the repoussé type, with subjects from classical mythology. Very few gold objects have survived, and most bronze work has also been lost. Decorative textiles of fine quality were common, and a few fragments have survived. It is in some of the rare fragments of patterned silks of the 7th or 8th century that the Persian influence is most often to be found. Silk at one time was imported in vast quantities from China.

Constantinople tended to become increasingly an Oriental city as the Greek influence introduced by Alexander the Great waned in the Near and Middle East and the new civilization of Islām was established. (Ge.S./Ed.)

Early medieval Europe. In the constant warfare that was waged in Europe in the early medieval period, material possessions dwindled to a minimum: a man did not own for long anything he could not defend and had little use or opportunity for interior decoration. If he possessed more than one house, his furniture and possessions would go with him from place to place. During this time, the arts came to be monopolized by the church, which grew to dominate all aspects of the medieval world.

By the 9th century the Romanesque style was well established in northern Europe. It made far greater use of the semicircular arch and vaulting than had the Imperial Roman style. Much of the sculpture decorating buildings was influenced by the Middle East. The court of Charlemagne in the 9th century was in communication with that of the caliph Hārūn ar-Rashīd, in Baghdad, and the Arabs had opened up a sea route between the Persian Gulf and China. Oriental textiles, imported through Venice and Genoa, began to be found in the more luxurious European interiors, and in the 13th century the first piece of Chinese porcelain, brought back by Marco Polo, found its way to the West and is still preserved in the treasury of St. Mark's, Venice.

Castle interiors

Late into the medieval period, the larger houses, generally called castles, were designed according to military rather than aesthetic principles. The main room was a spacious hall with timber or stone walls (sometimes plastered), an open-beamed roof, narrow slit windows (as yet unglazed), and a floor of stone slabs, tiles, or beaten earth. In the earlier houses the fire burned in the centre of the floor, and the smoke either drifted through a central hole in the roof or dispersed among the rafters; but wall fireplaces soon replaced this unsatisfactory system. Furniture was probably limited to plain stools, benches, and trestle tables, made of local timber, and some heavy chests in which personal possessions were stored. The feudal lord and his lady sat on more elaborate chairs on the dais (raised platform), and a coloured hanging of plain fabric sometimes decorated the wall behind them. Wall hangings and tapestries became more common in Norman times (1066–1189), when stone carving on doorways, fireplaces, window openings, column capitals, and arcading superimposed on the inside walls was also introduced. Such hangings can still be seen in the Norman castles of Rochester, Kent, and Chepstow in England. The whole community often lived and slept in the one hall, but as time went on, two main rooms—the hall and the chamber—were provided. At first, rooms were divided by woolen hangings, hung from iron rods or from the rafters. The houses of the poor were simple, timber-framed shelters with bare earth floors and undecorated walls. Such conditions, with variations according to local circumstances, were generally prevalent in western Europe until the end of the 12th century.

Late medieval Europe. During the 12th and 13th centuries those who had taken part in the Crusades learned something of luxurious living in the Near East, and as a more secure way of life was becoming possible at home, they began to improve their own living conditions. The castle slowly evolved into the manor house. Household equipment became more elaborate and important, no doubt partly because the women had played a greater part in household management since the absence of the men on the Crusades.

Curtains of finer texture began to replace wooden window shutters or heavy homespun hangings. Tapestries relieved the bareness of the walls and gave additional warmth to rooms, and other textiles and tapestries were draped over chairs and tables, and brightly coloured woven or embroidered cushions were used. The fine wood ceilings of the large rooms were sometimes coffered and often painted in bright colours, particularly in France. The disappearance of much of this colour with the passage of time lends a false austerity to surviving medieval interiors.

Development of specialized rooms

A greater number of rooms, serving special needs and giving increased privacy, came into use, although the house was still not planned as a whole. The kitchen, buttery, and pantry were placed at the lower end of the hall beyond a carved timber or stone screen, which, in larger houses, supported a minstrel's gallery. At the opposite end, there was a chamber, or withdrawing room, perhaps with a solar (upper room) above it, used as a bedroom or as a special apartment for the ladies. A guest room was occasionally provided. In the 13th and 14th centuries, the wardrobe was a room with presses for storing curtains, hangings, bed and table linen, as well as the clothing and materials needed by the members of the household.

Giraudon—Art Resource



Figure 19: Gothic courtly dining hall with tapestry covered walls, plaited rush mats, trestle table set with gold and silver tableware, and a side table for displaying household plate; Duc de Berry at table from the illuminated manuscript *Très Riches Heures du duc de Berry* by Pol de Limburg and his brothers, France, before 1416. In the Musee Condé, Chantilly, France.

Here sewing and tailoring were carried on, and the room became a combined workroom and storeroom, furnished with heavy, plain tables and chairs.

In the kitchen, rotating spits and adjustable hooks for suspending cooking pots were fixed into a vast hooded or recessed wall fireplace. Plain but pleasing utensils of wood, copper, and iron were kept on hooks on the walls, and enormously solid tables stood on the stone or tiled floor, which was strewn with sawdust or rushes (Figure 21). In the hall the rushes were mixed with fragrant herbs and helped to absorb some of the dirt, smells, and grease. By the 15th century plaited rush mats were common (Figure 19). The introduction of linen tablecloths resulted in a great improvement of manners and cleanliness at meals.

Ornaments and various luxuries, which had become more common during the time of the Crusades, proliferated in subsequent centuries as commerce with the Near East increased. Household plate, of gold or silver, was frequently displayed on dressers or cupboards as decoration and to impress visitors (Figure 19), and it was not unknown for these possessions to be roped off to prevent pilfering. Indoor arrangements for washing and bathing were considered a luxury. A flat-sided metal bowl was sometimes fixed to the wall of a living room with a swinging ewer or a small cistern with a tap over it and a towel on a hinged rod. Small convex mirrors were hung in the walls as early as the 15th century, such as the one in the background of Jan van Eyck's "The Marriage of Giovanni Arnolfini and Giovanna Cenami" (Figure 20).

The Gothic style

The Gothic style first made its appearance in the Ile de France, toward the end of the 12th century. It derived originally from Middle Eastern sources and was developed by Islāmic builders. It came to be widely employed in western Europe, where, for uncertain reasons, it gained the name Gothic by the 17th century. It is characterized by the extensive use of the pointed arch, by spacious interiors, and by walls pierced with numerous windows, often of stained glass (see Figure 1). The style had no fixed rules governing proportion, and decoration, generally, was the free expression of craftsmen within the limits of current fashion and the purpose of the building.

By courtesy of the trustees of the National Gallery, London; photograph, J.R. Freeman & Co. Ltd.



Figure 20: Northern Gothic bedroom with canopied bed, convex mirror, Oriental carpet, and brass chandelier, "The Marriage of Giovanni Arnolfini and Giovanna Cenami (?)," panel by Jan van Eyck, 1434. In the National Gallery, London.

Knowledge of Gothic interiors derives from illuminated manuscripts (Figure 19) and panel paintings (Figures 20 and 21) from the few surviving *objets d'art*. Much use was made of textiles for covering walls, especially tapestries (Figure 19); the principal medieval centres of tapestry manufacture were Paris and Arras (see the section *Tapestry* below). European courts at this time were very mobile and moved from place to place: tapestries were remarkably versatile, for they could be taken down and rehung elsewhere. They were employed to partition rooms, and were sometimes suspended under a high roof to act as a ceiling. Rugs and carpets had been brought back from the East by the crusaders and were at first employed as a covering for a divan or, in the case of the finer varieties, as bed and table coverings. The carpet for the floor was introduced comparatively late (Figure 20). Weavers of Saracen origin had settled in Sicily and on the Italian mainland, and they produced all kinds of rich fabrics, such as silk and velvet.

Use of hangings

Furniture was not present in such quantities as in later centuries, chairs especially being fairly rare. Tables were long and rectangular, laid on trestles, with benches for seating (Figure 19). At the head of the table, for the principal person of the household, was a straight-backed chair. Chairs, generally, were the subject of a certain etiquette, being reserved for the most important people, and they were often surmounted by canopies. Retainers had to stand (Figure 19); less important members of the household were sometimes supplied with stools. Folding chairs, like the old Roman curule chair, appeared in the 14th century. Although a few chairs had seats and arms stuffed with rushes, it was more common to drape them with textiles and put cushions on the seats. Buffets, often superbly carved, were used as a stand for silver and for serving food.

Medieval bedsteads, with highly carved posts and canopies, were often of great size, and they were customarily occupied by several persons—as well as the favourite dogs, who slept on top (Figure 20). The Great Bed of Ware in the Victoria and Albert Museum, London, is reputed to have held six couples in comfort.

Goldsmiths' work was often decorated with enamel, and bronze was similarly treated. The usual technique was the champlevé type, in which the metal is engraved or carved and the spaces then are filled with powdered coloured glass, subsequently fused by firing. At Limoges and in the Rhineland a wide range of objects were executed: quite large works, such as tombs, as well as smaller pieces, such as *chasses* and reliquaries. Lighting appliances were made of bronze or wrought iron. Those for suspension were usually intended for oil lamps, and standing candlesticks and candelabra were provided with spikes onto which the candle was forced (pricket candle sticks).

Gothic decorative pottery

Very little decorative pottery was made, although the colourful dishes and vases of Moorish Spain are an exception. Tiles were extensively employed for both walls and floors in houses of the better class, and there was a proverb in Spain to the effect that a poor man lived in a house without tiles (see Figure 30). The technique of manufacture was often quite complex and included inlaying with clay of a different colour. The vogue for tiles was imported from Islām by way of Moorish Spain. Chinese porcelain was known in western Europe by the late 14th century but was, of course, extremely rare; indeed, specimens were often mounted in silver in the same way as the semiprecious hard stones such as amethysts, garnets, and peridots.

The Gothic style lingered in England and northern Europe much longer than it did in the south, and many more examples of it escaped destructive wars than on the Continent. The panelled room characteristic of the style and the period has survived more or less intact in England, where panelling with traces of paint can still be found.

Gothic ornament sometimes makes use of motifs similar to those of classical interiors, such as the acanthus leaf and the rosette, but the treatment is very different. The Gothic craftsman liked to abstract certain features of his model and emphasize them in a stylized manner, as in the heraldic eagle, especially as it is used on the reverse of dishes from Moorish Spain and in coats of arms like that of the Holy

Roman emperor. It no longer bears any resemblance to the naturally depicted Roman eagle but is stylized, with a geometrically drawn tail. Similarly, the lion has its open mouth, tongue, mane, tail, and claws treated in the same way. Compass work is a marked feature of much Gothic ornament. The cross, for instance, is never a plain cross but is ornamented with geometric motifs; it may represent a reemergence of some old Celtic motifs, which were often based on compass work. Much Gothic ornament is floral and foliate, freely and naturally treated in some cases but stylized in others. Like interiors, paintings were in bright colours. Some of the ornamental motifs to be found in objects intended for interior furnishing are architectural, like the crocket (projections in foliate form), the panelling of chair backs, and the doors of buffets (Figure 21).

By courtesy of the Museo del Prado, Madrid



Figure 21: 15th-century Flemish interior with Gothic ornamented furniture: "St. Barbara," oil on panel, attributed to Robert Campin, 1438. In the Prado, Madrid.

Islāmic countries. The Arab conquest in the 7th century AD and, in the 8th century, Muslim expansion into India and Spain had profound influence on the decorative arts throughout the known world, especially as most of the long-distance trading routes passed through Arab lands. The skills of the conquerors fused with the traditional skills of their subject peoples, and because Islām forbade the portrayal of human or animal form, whether for religious or artistic purposes, and encouraged the incorporation of Qur'ānic texts into design, religion played a considerable and direct part in the development of design. As with nearly every other society, the finest and most lasting buildings were of a religious nature, and, unfortunately, few domestic dwellings have survived.

Architectural quality and form were subordinated to intricate and richly coloured surface decoration. Perhaps the finest results were achieved in Persia, where a high level of technical ability already existed in combination with great lyrical sensitivity. There the principal decorative features were the ceramic tiles and tile mosaics that encrusted floors, walls, roofs, and domes both inside and out (Figure 22). The mosques of Isfahan, Meshed, and Tabriz, ranging in date from the 13th to the early 17th centuries, demonstrate a completely satisfactory use of colour in architecture. Lustred tiles with a combination of floral and geometric design date from the 10th and 11th centuries, and naturalistic flowers frequently give a gardenlike effect to the tile decoration. Iris, rose, carnation, tulip, pomegranate, pine, and date are depicted, always with delicately interlacing stems, and contained within plain or patterned borders. Blues of all shades, from turquoise to a deep ultramarine, are characteristic.

Islāmic tiles and mosaics

Patterns for tilework and patterns for the Persian carpets are frequently interchangeable. Carpet designers soon managed to circumvent the Muslim ban on the use of animal forms: lions, deer, leopards, ornamental birds, and, occasionally, even mounted huntsmen were depicted, the figures always judiciously placed to give the maximum decorative effect. Artistic achievement reached its peak under Shah Abbās I (AD 1588–1629), but well before this time Persian carpets, silks, and pottery were known and valued among Europeans, as they still are in the 20th century.

In Egypt and Sicily one of the results of Muslim domination was the introduction of a high degree of ornamentation on wall surfaces, once again principally by means of vividly coloured ceramic tiles. The patterns are more solid than those of Persia, filling up the areas between the containing arabesques and with less open backgrounds. Moorish design in Spain shows even more complex interlacing geometrical framework, which is filled in with formalized leaves, flowers, or calligraphic inscriptions. Ceilings and the upper parts of walls were modelled in flat relief with coloured and gilded arabesques, while the lower wall areas were tiled. The decoration was partly hand chiselled and partly molded. Such decorations may be seen in the Alhambra, built at Granada in the 15th century, a pleasure palace whose arcaded courts and halls are embellished with stuccoed decoration in honeycombed ceilings, stalactite vaults and capitals, tiers of horseshoe-shaped or stalactite-fringed arches, and pierced or latticed windows.

Decoration in the Alhambra

In the mosques of Turkey, walls were veneered with marble, and ceramic tiling was introduced only in small areas. Colours, too, are less exuberant in the large mosques, where a sense of space rather than of overwhelming decoration is preeminent. Domestic buildings were largely of wood, looking inward to secluded courtyards and gardens, but with elaborately latticed windows projecting at upper-floor level over the street. As in most other Islāmic countries, the wealthy furnished their houses with velvet and silk hangings, couches, and innumerable cushions (Figure 22).

Islāmic influence in India appears at its finest in the interiors of mosques, tombs, and palaces built during the Mughal period (1556–1707).

Renaissance to the end of the 18th century. The Renaissance was a revival of the old classical styles, and it is not surprising that it first showed itself to a marked degree in Italy. The Gothic style had made comparatively little headway in Italy, where it was regarded as barbarous except in some of the more northerly towns, such as Milan and Venice. The style had more or less coincided with a period of primitive commerce. With the Renaissance the complex commercial organization of ancient Rome began to be revived by the towns of Tuscany, especially Florence. Feudalism disappeared, and the bourgeois merchants and financiers of the town rose to power and influence. Money began to circulate, banks were established, checks and bills were honoured over long distances, factories were opened, and men grew rich enough to buy and commission works of art for interior furnishing from those who owned their own workshops, employed assistants, and were no longer reliant on a system of patronage. With the rise of the town and the invention of gunpowder, the fortified country house became obsolete.

Social influences of design

In and around Florence the new commercial civilization was most highly organized. The old Greek and Roman manuscripts had been preserved, not only by the Christian monasteries but to an even greater extent by the Muslims, and soon after 1350 these began to find their way into northern Italy. Men became increasingly dissatisfied



Figure 22: Manuscript illumination depicting the intricately patterned geometric and floral ceramic tilework characteristically used in 16th-century Persian interiors. "Bahrām Gūr in the Yellow Pavilion on Sunday," illustration from the *Khamseh* of Nezāmī, Tabriz School, 1524–25. In the Metropolitan Museum of Art, New York City.
By courtesy of the Metropolitan Museum of Art, New York, gift of Alexander Smith Cochran, 1913

with the spiritual outlook of medieval Christianity, and the old Greek curiosity and philosophical speculation began to revive.

The Renaissance was, in fact, a return to the mainstream of Western art after what could fairly be described as the Gothic interregnum. Nevertheless, a thousand years lay between the fall of the Roman Empire and the Renaissance, and the classical styles of the Renaissance bear the same kind of resemblance to those of Rome as modern Italian bears to Latin. They are similar, but by no means the same thing.

The Renaissance brought back the Roman vocabulary of ornament, although the emphasis was now sometimes in different places (Figure 23). The classical orders (columns with base, shaft, capital, and entablature) were borrowed, and adapted to dress the new architectural style. Architects became highly skilled in the treatment of space, and decoration often played a major part in defining and enriching their vigorous spatial effects. Classical architectural forms were used in plasterwork, inlaid woodwork, and painted decoration as well as for staircases, doors, windows, and fireplaces, which formed increasingly important and elaborate features of interior design. Decorative details inspired by the antique were also used, executed in a wide variety of techniques; garlands, caryatids (statues of women used as supporting pillars), lion masks, grotesques, reclining amorini (cupids), cornucopia (horns overflowing with flowers or fruit), arabesques (entwining scroll and plant motifs), and trophies of arms are among the most familiar. Floors of coloured and patterned marble paving are frequently integrated with the overall decorative scheme. Modelled stucco, sgraffiti arabesques (made by cutting lines through a layer of plaster or stucco to reveal an underlayer), and fine wall painting were used in brilliant combinations in the early part of the 16th century.

In Venice the transition from Gothic to Renaissance building came less abruptly, as demonstrated in the Doge's Palace, where a Gothic exterior is found in combination with a late 15th-century facade on the east of the courtyard and a series of High Renaissance council chambers, famous for wall paintings by the Venetian painters Paolo Veronese and Tintoretto. Wood panelling with flat pilasters and a molded frieze forms the lower part of the interior wall decoration, with the fine series of historical and allegorical paintings, above, divided into panels between painted and gilded moldings and pilasters. The ceilings of a later date are particularly richly painted, their heavily scrolled carved and gilt cornices and framing introducing a touch of the Baroque style. Windows with twin semicircular headed frames surmounted by a lunette (a semicircular wall area) and fitted into a third, larger round-arched opening are a typically Venetian feature of the waterside palaces. In these, as in all the great Italian houses of the time, the works not only of the finest painters of the period but of the sculptors, goldsmiths and silversmiths, wood-carvers, bronzeworkers and ironworkers were used to embellish the principal rooms. Silks, embroideries, and cut velvets were used as hangings and upholstery, together with elaborately cut and framed looking glasses and carved gilt pendant chandeliers, as in the Palazzo Corner-Spinelli, Venice (1480). Costly carpets were imported, and much fine linen was in use. Trompe l'oeil (realistic) effects of perspective were achieved in the painting of walls and ceilings and also with intarsia (inlaid wood) decorated panelling such as in the study of Federico da Montefeltro, formerly at Gubbio, Italy and now in the Metropolitan Museum, New York City (Figure 24) or in the Palazzo Ducale, Urbino (completed about 1500), where a startling effect is created simulating open cupboards full of books.

Venetian interiors

During the Renaissance, Venice became a glass-making centre and introduced many new techniques. Blue glass with fine enamel painting dates from the end of the 15th century. Excellent engraving was done with a diamond point as soon as glass of sufficiently good colour was produced, by using manganese to neutralize the colour introduced by impurities in the raw materials. Such glass, which was called *cristallo* from its fancied resemblance to the hardstone known as rock crystal, is the origin of modern crystal glass. The Venetians also imitated coloured hardstones in glass. Glass made white and opaque with tin oxide was sometimes used for enamel painting in the style of porcelain, and clear glass with opaque white threads embedded in it in lace-work patterns was called *vitro di trina*. The Venetians also made mirror glass of excellent quality; in the 17th century they supplied the mirrors for the Galerie des Glaces of the palace of Versailles. Large sheets, however, were not practicable until the French discovered a method of making plate glass late in the 17th century, when the national factory of Saint-Gobain was founded.

During medieval times, Italian wood-carvers had achieved a high level of skill in the decoration of churches; now they turned to secular furniture, for which they employed oak, walnut, cypress, and a new, rare, and expensive wood—ebony. (In 17th century France, the craftsmen skillful enough to be entrusted with this wood—who were also makers of cabinets—came to be called *ébénistes*, a term that remains the French equivalent of the English "cabinetmaker.") Many ancient Roman furniture-decorating techniques were revived. Inlaying with a variety of coloured woods, with ivory, mother-of-pearl, and tortoiseshell, with a mosaic of coloured stones known as *pietra dura*, and with painting and gilding in addition, ornamented the finest furniture. The chest (cassone), often commissioned on the occasion of a wedding, was decorated with elaborate painting and gilding, sometimes with a large pictorial subject and sometimes with elaborately carved work, which was later coloured. Italian furniture in its design often made use of architectural motifs. Cabinets were often exceptionally luxurious, with such elements as caryatids flanking central doors, arcades of semicircular arches, and triangular pediment tops. The interiors were sometimes small models of architectural interiors, with

Italian wood-working

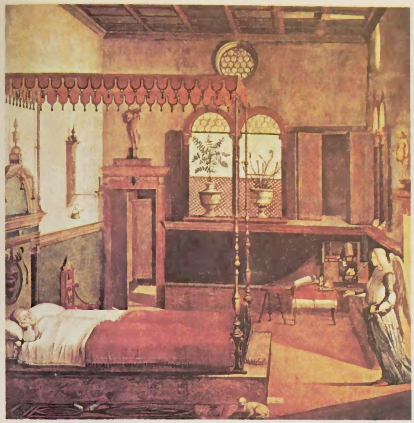


Figure 23: Classical ornament used in Italian Renaissance interiors, "Dream of St. Ursula," canvas by Vittore Carpaccio, Italy, *c.* 1495. In the Accademia, Venice.
SCALA—Art Resource

mirrors inset to give an impression of spaciousness. Silver furniture, no longer extant, was used in considerable quantities in late Renaissance times, usually crafted from plates of silver beaten over wooden formers.

Majolica

An innovation in Italy, which rapidly spread throughout the rest of Europe, was tin enamelled pottery, known in Italy as majolica and farther north as faïence or delft. Colourful dishes were often painted in a style known as *istoriato* (history painting) with mythological and biblical subjects. As some of the subjects were taken from engravings of Raphael's work, this pottery became known during the 18th and 19th centuries as Raffaelle ware. The majolica potters, the best of them located in Tuscany, made extensive use of grotesques, which show the style at its best.

The old Roman fashion for small bronze figures was revived during the Renaissance, and the fashion for these in interior decoration continued almost to the end of the 19th century. The earliest were fairly exact copies of excavated classical bronzes and may have been forgeries intended for sale at the time as genuine Roman work. The art developed rapidly. Before the 16th century, bronzework was done by the goldsmiths, and, as in most goldsmiths' work, general effect was subordinated to meticulous detail. After 1500, when bronze became popular for lamps, candlesticks, sconces, inkstands, small freestanding decorative figures, and furniture mounts, treatment of suitable subjects developed along the lines laid down for full-sized sculpture. Many small bronzes were made, some of them in the grotesque style.

At the beginning of the 16th century, the revived classicism of the Renaissance began to be modified, and eventually the style divided into two distinct paths. One remained faithful to tradition. The architect Andrea Palladio took ancient Roman works as a model, basing his designs on the theory of proportion laid down by Vitruvius in the 1st century BC in the *Ten Books on Architecture.* The second path was initiated by Michelangelo and led by way of Mannerism to the Baroque style. In both these latter styles, a deliberate exaggeration of forms displaced the strict logic and precision of the High Renaissance and aimed to convey freedom of movement and to involve the spectator in the drama of the design. Mannerism had only a limited influence on interior furnishing, as in the bronzes by Cellini and by Giambologna. Poses are often strained, the torso twisted, and the musculature emphasized; the favourite Mannerist subjects are violent ones, such as the rape of the Sabines and Hercules slaying Antaeus.

Characteristics of the Baroque

Baroque was the style of the Counter-Reformation and was intended by the Jesuits to express the temporal power and riches of the Catholic Church in contrast to the austere doctrines of Protestantism. The theatricality of the Baroque style soon attracted the attention of princes, who wanted it to be used in the palaces they built (Figure 25). Coloured marbles were used extensively, frequently in combination with bronze and rich gilding. Coloured glass windows were often used for lighting special features. Walls were sometimes painted to appear to be a continuation of the interior, giving an impression of spaciousness. Certain materials were often simulated by others: *scagliola,* for example, is a mixture of marble chippings, gypsum, and glue that was widely employed to imitate brecciated marble. What appeared to be richly coloured marbles were often no more than painted wood. Drapery was frequently imitated in carved marble, and wooden columns, the purpose of which was purely decorative, were painted like marble or some other exotic stone. Marble or stucco was made to imitate brocaded hangings, as in the Sala Ducale, Vatican, where an effect of space from limited means is created. Basic techniques were unaltered, but all restraint in their use vanished in bold theatrical effects and sensual luxuriance of modelling. Walls became curved, pediments were broken (*i.e.,* with central part omitted), columns and pilasters twisted until the buildings seem to come alive with movement. Bernini exuberantly combined rockwork, figures, and draperies with columns, panelling, and vaulting.

From Italy these styles spread across Europe, where they were absorbed in varying degrees and tempered by the national or local taste and genius. Many Italian designers and craftsmen travelled and worked abroad in France, England, Austria, and Spain.

France. From the middle of the 15th century, ideas from Italy began to change the face of French buildings; this change came gradually, first in the applied decorative detail superimposed on basically Gothic designs, then extending to a symmetry and regularity of the whole. Indeed, one of the basic differences between the Renaissance in France and in Italy is that in the latter the revolution in style involved, from the very outset, the whole conception

By courtesy of the Metropolitan Museum of Art, New York, Rogers Fund, 1939

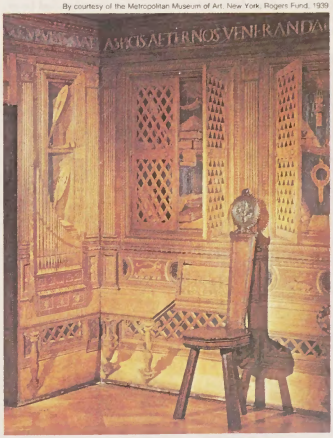


Figure 24: Intarsia panels from the small study of Federico da Montefeltro, duke of Urbino, at Gubbio, attributed to Francesco di Giorgio of Siena, *c.* 1480. In the Metropolitan Museum of Art, New York City.

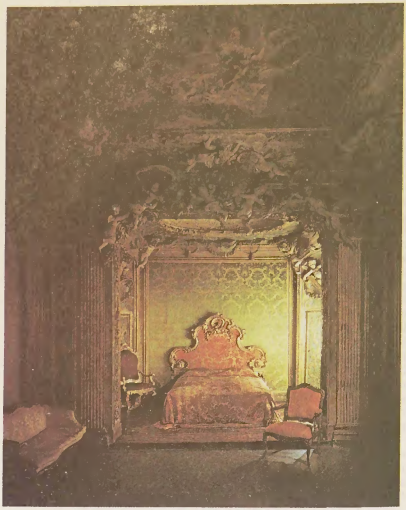


Figure 25: Baroque cupids supporting painted and gilded ceilings in the theatrically conceived bedroom of the Palazzo Sagredo, Venice, c. 1718. In the Metropolitan Museum of Art, New York City.

By courtesy of the Metropolitan Museum of Art, New York, Rogers Fund, 1906

of design. The centralization of power and the brilliance of French court life was consolidated under Francis I (1515–47) and had already resulted in patronage of artists and craftsmen from Italy. Since the need for churches had been fulfilled in the great age of Gothic building, the king and his court rivalled one another's magnificence in building new châteaux in the early Renaissance style. Stone and timber were readily available, with masons and carpenters skilled in their use.

Style of Francis I

Among the earliest attempts in the new manner are the additions made by Francis I to the Château de Blois. The spiral staircase, with its own open stonework tower, may have been designed by Leonardo da Vinci, who died nearby at Amboise in 1519. Even at this early stage, the decoration of the staircase ceiling with carved bosses (an ornamental ceiling projection) featuring the monogram and heraldic device of the king shows a typical French contribution to Renaissance decoration. Such shields and monograms formed an important element in many decorative features, being used in wall and ceiling panel design or on the large carved stone chimneypieces. The fine galleries of Francis I and Henry II (1547–59) in the royal Palais de Fontainebleau illustrate the increasing elaboration of applied decoration and colour (Figure 26). The flat ceilings are of wood, coffered, coloured, and gilded in a variety of geometrical forms outlined with fine moldings. Molded panels enclose paintings on the upper section of the walls, and molded or carved wood panelling the lower parts, as in Italy. Floors are of hardwood strips, sometimes repeating the pattern of the coffered ceiling above. Benches supported on consoles (ornamental brackets) are designed as part of the overall scheme of wall panelling. Italian artists had been employed at Fontainebleau and elsewhere, influencing the contemporary French architects toward a more Italian conception. Rosso Fiorentino and Francesco Primaticcio decorated the Galerie de François I, while the hexagonal coffered ceiling in the Galerie de Henri II was designed by the French architect Philibert de l'Orme. The architects Sebastiano Serlio and Giacomo da Vignola, together with the goldsmith Benevenuto Cellini, all worked for a time in France, and much of the decorative work in the châteaux of the Loire valley was executed by Italian craftsmen.

In the early 17th century and during the long reign of Louis XIV (1643–1715), formality and magnificence became paramount in the life of the court. Suites of large rooms elaborately decorated provided an opulent background for the King and his courtiers; such suites usually consisted of a vestibule, antechamber, dining room, salon, state bedroom, study, and gallery. Staircases were stately and spacious, offering a fitting approach to the main rooms. Decorative schemes incorporated the fittings, hangings, and furniture with that of the room.

Giraudon—Art Resource



Figure 26: Elaborately carved and painted gallery characteristic of French Renaissance design: Palais de Fontainebleau, Galerie de François I, c. 1533–45.

Versailles

The Baroque style was admirably fitted to express ideas of luxury and pomp. It inspired the building of some of the finest palaces erected in Europe since the days of Imperial Rome. The palace of Versailles built in the mid-17th century and widely imitated, led to the French court style in interior decoration and furnishings. Versailles was intended to be the outward and visible expression of the glory of France, and of Louis XIV, then Europe's most powerful monarch. His finance minister, Colbert, set up a manufactory that made works of art of all kinds, from furniture to jewellery, for interior decoration. A large export trade took French styles to almost every corner of Europe, made France a centre for luxuries, and gave to Paris an influence that has lasted till the present day. The vast initial cost of Versailles has been more than recouped since its completion. Even Louis XIV's most violent enemies imitated the decoration of his palace at Versailles. In 1667 Charles Le Brun was appointed director of the Gobelins factory, which had been bought by the King, and Le Brun himself prepared designs for various objects, from the painted ceilings of the Galerie des Glaces (Hall of Mirrors) at Versailles to the metal hardware for a door lock. (It should be noted that at the Gobelins, as elsewhere in France, furniture was designed by artists or architects who had no practical experience of manufacture, whereas, in the great age of furniture making in England, most designs were made and executed by the cabinetmaker himself, who had an intimate knowledge of his material.)

Though the Baroque trend is well established in the Versailles interiors, generally speaking it was regulated in France by an underlying restraint that seldom permitted decoration or movement to dominate entirely. Besides the Galerie des Glaces at Versailles (Figure 27), the Galerie d'Apollon at the Louvre is an example of magnificence in decoration. The vastness of these rooms and the lavish use of marble, plasterwork, and painted ceilings (with the addition at Versailles of mirror glass panels) created an effect of overwhelming grandeur.

Giraudon—Art Resource



Figure 27: Formality and magnificence appropriate to the court of Louis XIV: Galerie des Glaces (Hall of Mirrors), Versailles, designed by Jules Hardouin-Mansart, ceiling painted by Charles Le Brun, 1678.



Figure 28: A delicacy of decorative motif in panelling and furniture characteristic of the Rococo design of the Louis XV style: room from the Hôtel de Varengeville, Paris, design attributed to Nicolas Pineau, c. 1735. In the Metropolitan Museum of Art, New York City.

By courtesy of the Metropolitan Museum of Art, New York, acquired with funds given by Mr. and Mrs. Charles B. Wrightsman

Transition from Baroque to Rococo

Among the architects and artists working at this time were Jean Berain, André-Charles Boulle, Jean le Paultre, Robert de Cotte, and Jules Hardouin-Mansart. Their work continued in the later period in which Baroque ornament was transformed into the airy, delicate Rococo of the mid-18th century (Figure 28). The beginning of this more fluent treatment can be seen in the work of Robert de Cotte at Versailles and the Hôtel de Toulouse, Paris. An immense variety of materials was used for the inlaid and decorated furniture; in a piece by Boulle, for instance, the designer employed—in addition to the tortoise-shell and brass inlay—ebony, copper, lapis lazuli, green-stained ivory or horn, and mother-of-pearl.

Despite its freedom from onerous restrictions, the Baroque style had preserved the classical idea of symmetry. Not until the early decades of the 18th century were there marked departures from the notion that an object divided vertically should consist of two halves that are mirror images of each other. The Louis XIV style embodied a passion for symmetry, but with the Regency of the duc d'Orléans, which began in 1715, asymmetry became one of the features of contemporary decoration and one of the major aspects of the Rococo style. The principal designer in this style, who was largely responsible for its development, was Juste-Aurèle Meissonnier, a goldsmith and *ornemeniste*. It is no accident that many objects in Rococo style, including furniture, look as though they had been designed by a metalworker. It has been said that Rococo began when the scrolls stopped being symmetrical. The influences that brought about this revolutionary concept are worthy of consideration.

The influence of oriental porcelain

Beginning in the early decades of the 17th century, Chinese porcelain and lacquer were imported into Europe in ever-increasing quantities. Porcelain, especially, attracted many distinguished collectors, including most of the royalty of Europe. This increasing use of Chinese art objects in European decorative art provided a powerful influence with no trace of classical tradition. Soon after 1650 the Dutch began to import porcelain from Japan, at first decorated in blue, but toward the end of the century in polychrome, painted either by, or in the manner of,

Sakaida Kakiemon. This was widely sought, and even more highly valued than Chinese porcelain. When Augustus the Strong, elector of Saxony and king of Poland, bought a palace early in the 18th century to house his collection, for instance, he called it the Japanische Palais, and in France Louis-Henri de Bourbon-Condé, duc de Bourbon, established a factory at Chantilly to imitate Japanese porcelain. The decorations of Kakiemon were markedly asymmetrical, as were the painted lacquer panels that were imported to be made into screens and furniture, and there seems no doubt that this feature also influenced European Rococo art.

Despite the quantities in which it was imported, the demand for Oriental porcelain could not be satisfied, and European potters sought desperately to discover the secret. The first factory to make porcelain in the Oriental manner was at Meissen in Saxony, patronized by Augustus the Strong, but soon many small factories began to spring up in Germany, Austria, and Italy. France had several factories making a modified type of porcelain, the most important being the Sèvres factory, owned by Louis XV and patronized by his mistress, Mme de Pompadour. The first English factory, at Chelsea, was established as late as 1745. Porcelain was probably the most important expression of the Rococo style in the first half of the 18th century, with bronze and goldsmiths' work closely following in second place; indeed, this period might well be called the age of porcelain. Rooms entirely decorated with porcelain still exist. These included not only vases and figures, but also mirror-frames, scrollwork, cornices, and even small console tables. A very fine example still survives at the Palazzo di Capodimonte (Museo e Gallerie Nazionale di Capodimonte) in Naples.

The French style developed, in the 18th century, into a very skillful synthesis of materials in which bronze and porcelain played an important part. Furniture was elaborately mounted in bronze with a marble top and was often decorated with porcelain plaques, as well. Clocks were made from porcelain vases. Jardinieres and vases were filled with porcelain flowers with bronze stalks and leaves. Veneering with rare woods reached its height, and decorative marquetry, often elaborately pictorial, was practiced. Much sought at this time was the marquetry of brass and tortoiseshell, which began with Boulle, although it was a revival of an Imperial Roman fashion. Tapestries covered the walls when these were not decorated with carved wood-panelling known as boiserie. Another form of wall-decoration, also employed in the making of furniture, was vernis Martin (Martin's varnish), an imitation of Oriental lacquer that was extremely popular after 1730. The large *salon de reception* of the 17th century gave place to smaller, more intimate rooms, and more of them, and the furniture and decoration of the period are also on a smaller scale.

The Rococo style is remarkable for its flowers and its curves. Furniture legs were gracefully curved, and tops were cut into serpentine shapes. It is easy to see when the Rococo style ends, because chairlegs at once become straight.

Typical Rococo features are seen in the interiors of the architect and decorator Germain Boffrand for the Hôtel de Soubise, Paris (begun 1732), where architectural form has been subordinated to the demands of the decoration; the cornice has disappeared, and walls curve into the ceiling, appliquéd with ragged C scrolls, garlands of flowers decked with ribbons, sprays of foliage, trellising, and shell motifs. The reduced scale of rooms and the reaction from monumental design result in elimination of the classical orders. Relatively small painted panels, idealizing peasant life, were enclosed in flattened moldings, silvered or gilt; pastel-coloured backgrounds prevented the smaller size of the salons from becoming too evident. The use of Chinese motifs typifies the search for novelty and blends well with the general lightness of style. The Cabinet de la Pendule (Room of the Clock) at Versailles (1738), designed by J. Verberckst, is another excellent example of French Rococo interior design. Gilles-Marie Oppenordt and François de Cuvilliés also were distinguished designers who worked with the best artists and craftsmen of the time.

The Rococo fashion spread across Europe to the courts of minor royalties, where many Frenchmen were employed to provide up-to-date buildings and schemes of decoration. In France the Gobelins factory became restricted mainly to the output of tapestries; equally fine work is seen in Aubusson and Beauvais carpets and tapestry. Improvement in glass manufacture resulted in larger mirror panels and brilliant crystal chandeliers.

Neoclassicism of the Louis XVI style

The Louis XVI, or Neoclassical, style began, in fact, to take root before the death of Louis XV in 1774; Mme de Pompadour and her brother, the Marquis de Marigny, were among the first to be attracted by the new classical style in the 1750s. From 1748 onward the characteristically French regard for formality was stimulated by the archaeological discoveries at the sites of the ancient Roman cities of Herculaneum and Pompeii and by the other surveys of classical remains published at this time.

It is sometimes forgotten that contemporary English styles also had influence in France, mainly through the published works of the architects Robert and James Adam. The asymmetrical, sinuous lines of the Rococo were slowly replaced by a more restrained form of decoration based once again on straight lines, right angles, circles, and ovals, arranged symmetrically. The lightness and fine moldings were retained, but the decorative forms were once more contained by the architectural framework. New motifs, many of them selected from antique Roman wall painting, decorated the panelling, in paint or in flat relief; palmettes, husks, urns, tripod stands, sphinxes, trophies of arms or musical instruments were frequently combined in the decorative schemes (Figure 29). Gilt bronze was used with wood and plasterwork for moldings and ornamental

By courtesy of the Victoria and Albert Museum, London; photograph, John Webb

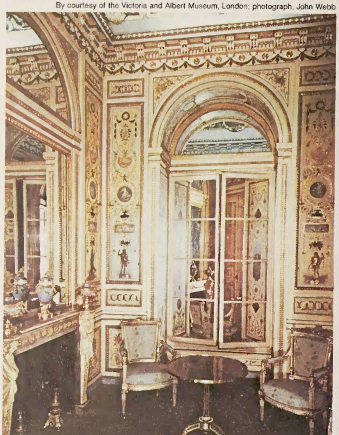


Figure 29: Symmetrical, restrained motifs based on the antique designs characteristic of the early Neoclassical Louis XVI style: boudoir of Madame de Serilly, Hôtel de Soubise, Paris, c. 1732. In the Victoria and Albert Museum, London.

fillets, emphasizing the rectilinear character of the design. The work of J.-A. Gabriel in both the Chambre du Conseil at the École Militaire (begun 1751) and the Galerie Dorée, Ministère de Marine (begun 1762) may be cited as Parisian examples. The keynote of colouring, as well as design, is refined simplicity. Silk tapestry wall hangings with fine flower and ribbon motifs appear in pale blues, greens, rose, and lilac. Similar colourings were used for satin and velvet upholstery. The fine wood carving of the brothers Rousseau, gilt bronze work by Clodion (Claude Michel),

and furniture pieces by David Röntgen, C.E. Riesener, and Jean Oeben show Louis XVI decoration at its highest. Apartments for Queen Marie-Antoinette at Versailles and her boudoir at Fontainebleau are full of this extravagant delicacy, soon to be obliterated in the French Revolution.

Spain. In Spain, Moorish influence mingled with subsequent Western classical styles to produce a unique flavour in decorative design. The style known as Mudéjar (*c.* 12th–17th century) was the early outcome of these blended Christian and Arab ideas and consists in essence of tiled floors and skirtings in polychrome (Figure 30), plain white walls, carved stucco friezes, and intricately decorated beamed wooden ceilings. The Duke of Alba's palace, Seville, contains fine interiors decorated in this style.

The Mudéjar style

Yellow tiles decorated with freehand motifs in blue became common in the 16th century. Tiles were often used on the ground floor of summer living rooms. Since fireplaces were seldom used in southern Spain, these rooms were vacated in the winter for the upper rooms.

The discovery of the New World, with the riches Spain subsequently drew from Mexico and Peru, created a period of Spanish ascendancy in the 16th century that encouraged building and coincided with the spread of Renaissance ideas throughout Europe. The influence of decorative craftsmen from Italy, together with the abundance of precious metal, encouraged the development of Plateresque ("silversmith-like") decoration. This type of Renaissance decoration was first seen in church interiors, in the form of tombs, *retablos* (a decorative structure behind an altar), and ironwork screens. The Italian motifs were used in a totally non-Italian manner, encrusting the surfaces as in the late Gothic or Mudéjar style.

This unique Spanish blend of widely separate styles produced the fine interiors of the late 15th-century Panteón de los Duques del Infantado, Guadalajara, by Vazquez, and the Palacio de Peñaranda de Duero (*c.* 1530), probably by Francisco de Colonia, where interlaced ceiling beams and timber panels were supported on honeycomb cornices and finely ornamented friezes. (Unfortunately, much of this work is now damaged or destroyed.)

Smaller houses as well as palaces were built around a patio, usually colonnaded and with modelled or carved friezes, columns, and bracket capitals.

Window grilles, or *rejas*, often form an important part of the decorative scheme, the ironwork being traditionally of a high degree of excellence. Love of closely patterned decoration, enveloping all surfaces that could easily be carved or modelled, is an important characteristic of early Renaissance work in Spain, and of the contemporary Manueline style in Portugal. Similar, if rather coarser, work in this style flourished in the American colonies.

High Renaissance decoration in Spain was influenced deeply by the austere character of Philip II and his vast combined palace and monastery, El Escorial (1559–84), near Madrid (Figure 30). This was built for him by Juan Bautista de Toledo and Juan de Herrera. Much of the granite of which the monastery is built is left unadorned, and frescoed vaulted ceilings are the main decorative features of interior design.

A revival of decorative arts took place in the late 17th century under the influence of José Benito Churriguera, his family and followers. The Churrigueresque, which also remained a peculiarly Spanish style, expressed the Baroque feeling of the 17th century in extravagant polychrome. Surfaces were broken into scrolls, rosettes, volutes, and fantasticated moldings; bunches of fruit and flowers hung from broken or inverted cornice moldings; and the whole interior—for example, the Sacristy of the Cartuja, Granada (1727–64)—appears to drip with ornament. Here, even cupboards and doors were inlaid with silver, tortoiseshell, and ivory, and the only plain surface is the checkerboard tiled floor. Remarkable among domestic examples of this style is the Palacio del Marqués de Dos Aguas, Valencia (1740–44).

The Churrigueresque style

Under the Bourbons, French and Italian influence increased, as can be seen in the interiors of the Royal Palace at Madrid (1738–64), with its handsomely painted ceilings and brocade wall hangings. Here, also, subsequent changes of taste are echoed in the lighter Rococo treatment of the Gasparini Saloon. Toward the end of the 18th century the Neoclassical movement gained a limited footing, though regional styles continued to incorporate the Baroque and older forms.



Figure 30: Austere Spanish interior of the Renaissance period; apartments of Philip II: El Escorial, near Madrid, second half of the 16th century.
By courtesy of the Newsweek Book Division, photograph, Michael Holford

Fine examples of Spanish colonial work exist in Mexico, Peru, and other South American countries where the Baroque was allied, as in Europe, with the Jesuits. Churches are painted and gilded with an exuberance equal to or even greater than that found in the mother country. Sometimes the churches are encrusted with tiles, and they always possess elaborate *retablos*.

Northern Europe. After spreading from Italy to France, Renaissance influence began to filter to Belgium and Holland, later reaching the various Germanic states and finally dying out in Scandinavia and Russia.

In the Low Countries and northern Germany during the 16th-century Renaissance, ornament was adapted to form an entirely individual style, which can be seen in the pattern books of the artists Hans Vredeman de Vries and Wendel Dietterlin. Strapwork (interlacing bands) and raised faceted ornament were widely employed, together with muscular, grotesque masked caryatids and distorted architectural features arranged in undisciplined designs (Figure 31). Chimney pieces, with overmantels carried to the ceiling, were embellished with marble columns and elaborate strapwork patterns, while similar ornaments flanked the doorways and enriched the ceilings. The great tapestries for which the Netherlands had long been famous were still in use, and Oriental carpets were spread as table covers and not used on the floors. Many town houses and civic buildings were comfortably appointed, yet without spectacular extravagances, and give an impression of modest prosperity. In Belgium the Musée Plantin-Moretus, Antwerp (1550), is unusually richly decorated, showing the influence of Spanish rule in the use of embossed leather as a wall covering. Large windows, with rectangular leaded lights, are again typical of a northern climate. Ceilings are beamed or plastered, and floors most frequently are of tiles on the ground floor and timber on upper floors.

Use of strapwork

The later styles of Baroque and the 18th-century tastes are copied from French models, particularly in Belgium. The Dutch, after achieving independence in the latter part of the 16th century, developed their decoration on more individual lines. Typical domestic interiors on a small scale are familiar through the paintings of the 17th-century artists Jan Vermeer and Pieter de Hooch (Figure 32). The fine series of town houses by Daniel Marot and his sons in The Hague illustrate the cross-currents of the various styles; built at the turn of the 17th century, they were conceived in the Louis XIV, or Régence, manner, yet could

Marot town houses



Figure 31: Strapwork and facetted ornament: Swiss Renaissance room from the Rosenburg at Stans, Switzerland, 1602.
By courtesy of the Musee National Suisse, Zurich

be set down in 18th-century London without incongruity (and Marot did, in fact, work for a time in England). Fine stuccoed ceilings and overdoors, largely uncoloured, and wrought iron balustrading are characteristic.

In Germany the general trend was similar, but in southern Germany and Austria fresh impetus and individuality were given to Baroque and also Rococo design. French and Italian craftsmen worked throughout the 17th century on the many Catholic churches built in south Germany, Austria, Bohemia, and Moravia. The use of colour, fresco, and stucco that they introduced has its own particular flavour when seen in cool northern light (see Figure 12, left).

Secular building from the early 18th century, in the hands of such architects as Johann Lucas von Hildebrandt and Johann Bernhard Fischer von Erlach, makes use of much sculptural detail. Windows are round or oval, figures strain to support capitals, balustrades are carved in sculptural manner, and modelled niches contain larger than life-sized figures; all these give a feeling of movement reminiscent in its impact of Bernini's work in Rome. Another characteristic was the enormous staircase hall, or *Treppenhaus*, which was one of the most notable interior features of German and Austrian Baroque and Rococo architecture. In the halls, colour was frequently confined to the painted ceilings, giving increased force to the novel and delicious colours of the rooms beyond. A vermilion dado or olive-green panels may be contrasted with white and gold. In the Nymphenburg Palace, near Munich (1734–39), by the Frenchman François de Cuvilliés, the Rococo reaches its crowning achievement: mirrors are framed in freely scrolled moldings, which in their turn are interspersed with trellising, garlands, baskets of fruit and flowers, cupids, birds, and fountains in silvered stucco on a pale blue or yellow ground, the whole evoking the essence of pastoral Romanticism (Figure 33).

Scandinavian interiors

Mingled influences from France, Holland, and England reached Sweden and Denmark in the mid-17th century and are seen in the Baroque and Louis XIV interiors of the Riddarhuset and Royal Palace in Stockholm and in the chinoiserie (Chinese-influenced decoration) of the Royal Palace of Drottningholm. Scandinavian interiors, however, largely continued to be of the traditional exposed timber boarding, hung perhaps with painted linen panels and brightened by woven chair and cushion coverings (Figure 34).

Russia imported foreign designers and styles in the late 17th and 18th centuries for the palaces built under the westernizing influence of Peter I the Great, his daughter Elizabeth, and Catherine II the Great. In the mid-18th century the Italian Bartolomeo Francesco Rastrelli designed the Tsarskoye (Detskoye) Selo (now called Pushkin), Peterhof, and Winter palaces in or near St. Petersburg, and A.B. Kvasov, S.I. Chevakinsky, and Rastrelli designed the Hermitage, also in St. Petersburg. Each worked largely according to his own current national styles. The same is true of the work of the British architect Charles Cameron at Tsarskoye Selo Palace and Pavlovsk Palace.

In many areas of Europe, Renaissance, Baroque, and Rococo had little effect on interior decoration. In the Alpine lands, where wood was cheap and plentiful, traditional medieval methods continued for a long time. Wooden floors and ceilings and panelled walls, or partly panelled with plain plaster above, were the general rule. The moldings were bold, but carving was usually in low relief and often the woodwork was painted in bright colours.

England. The breakup of the feudal system during the Wars of the Roses and under Henry VII in the late 15th century had far-reaching effects on the social structure of the time and consequently on domestic buildings and their decoration. The new conditions necessitated a larger number of rooms, and a great hall, though still an important apartment, was no longer the focus of indoor life. Wider distribution of wealth gave rise to numerous coun-

By courtesy of the Rijksmuseum, Amsterdam



Figure 32: Domestic interior typical of the 17th-century Dutch home; "Maternal Duty," canvas by Pieter de Hooch (1629–c. 1683). In the Rijksmuseum, Amsterdam.



Figure 33: Sinuous, intricate curves characteristic of the Rococo decorative vocabulary: circular mirror room in the Amalienburg pavilion, Nymphenburg Palace, near Munich, designed by François de Cuvilliés, 1734–39.
Gunther Schmidt—EB Inc

try houses, and for the next 400 years the English excelled in their building and decoration.

Adoption of Italian ornament

The Italian style reached England in the early 16th century; the earliest example is the tomb of Henry VII in Westminster Abbey, designed by Pietro Torrigiani of Florence at the command of Henry VIII and completed in 1518. For the next 40 years or so, English craftsmen borrowed from the repertoire of Italian ornament, at first inspired by and imitating the Italian artists and craftsmen employed on royal works at Hampton Court Palace, Middlesex, and the Palace of Westminster, London, who used arabesque decoration, medallion heads, and amorini on panelling and plasterwork, often mingling them with the traditional Gothic motifs. The great hall at Hampton Court (1515–30) shows a combination of Renaissance carved and gilded detail with the traditional type of open timber roof, known as the hammerbeam roof, and windows divided into sections by vertical posts (mullions). In spite of Henry VIII's example, however, the Gothic style died hard in England, lingering in the remoter districts well into the 17th century.

During the second half of the 16th century, as a result of the break with Rome, the Italian style was largely replaced by the distinctive Renaissance style of the Low Countries and Germany, fostered by the close religious, political, and economic relations between England and the Low Countries, the influx of immigrant workmen, and the circulation of Flemish and German pattern books. This new manner became the dominant influence in the decoration of panelling and plasterwork, characteristic features being intrinsic strapwork patterns, pyramid finials (sculptured ornaments used to terminate roof gables), raised faceted ornament, masks and caryatid figures, scrolls, and pilasters. Both the Italian and Flemish styles were adapted and naturalized to some extent by the English craftsmen, producing a new style that is peculiarly English.

At this time, also, the internal porch was introduced into many houses; this device excluded drafts from the room and also in some cases made it possible to reach a second room without passing through the first.

The frescoing of walls continued; of the few remaining examples, some show scenes from biblical and classical sources and incidents from local folklore. A good Elizabethan example depicting scenes from the story of Tobit was found at the White Swan inn at Stratford-on-Avon. Embossed, painted, and gilt leather was less used in England than on the Continent, but tapestries and such woven fabrics as velvet and damask for the wealthy and "says" (fabrics resembling serge) and "bayes" (baize) for people of more modest means were widely used as wall coverings. The inventories of Henry VIII's palaces show the vast number of tapestries and various hangings possessed by kings and great men. Hangings of painted cloth were widely used as a cheaper substitute for tapestry; these, too, depicted incidents from biblical and classical sources and employed decorative motifs ranging from Gothic to Renaissance subjects. Nearly all of this "counterfeit arras" has perished. The plaited rush matting continued to be used as a floor covering in Elizabethan interiors.

Tudor wall decoration

Great chambers and long galleries, usually on the upper floors, are distinctively Elizabethan or Tudor and were used in many cases for work and recreation in bad weather (Figure 35). Barrel-vaulted ceilings occupying the roof space often increased the height of the rooms, as at Chastleton House, Oxfordshire (*c.* 1603). The plaster ceilings were treated elaborately; narrow interlaced bands formed geometrical patterns, with semistylized floral, arabesque, or heraldic motifs in the panels between.

The steep medieval winding newel stair (stair with central pillar from which steps radiate) in wood or, more often, stone was abandoned for the more spacious staircase with straight flights of stairs, easier in gradient and planned round an open well. This was most frequently constructed of oak, with carved newel posts (the upright terminating a flight of stairs) and balusters (individual columns in a balustrade) making the most of the opportunity offered for decoration and enrichment.

Staircase design

Toward the middle of the 16th century, a feeling for classic reserve was spreading and the late Renaissance period might have flowered under Charles I had not political upheaval checked the zest for fine building. The architect and stage designer Inigo Jones twice visited Italy and was one of the few north European architects completely to absorb the spirit and decorative repertoire of Italian Renaissance classicism. He introduced the new style in the Banqueting House at Whitehall, the Queen's House at Greenwich; and with his associate and kinsman, John Webb, built Wilton House, Wiltshire.

At Wilton the Double Cube Room (*c.* 1649) shows the nobility of effect Jones was able to achieve in a small compass, for the dimensions of the room—60 by 30 by 30 feet (18 by 9 by 9 metres)—are not large, comparatively speaking (Figure 36). The basic influence

By courtesy of the Nordiska Museet, Stockholm

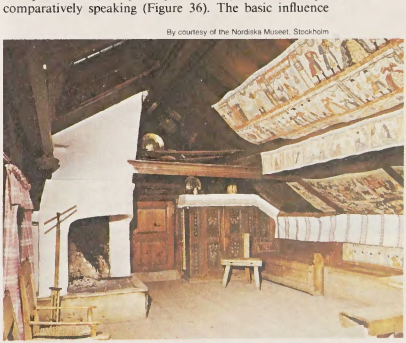


Figure 34: Pine panelled bedroom with painted linen hangings, Oktorp farmstead, Stockholm, 18th century. In the Skansen, Stockholm.



Figure 35: Panelled walls, tapestries, and intricately molded plaster ceilings characteristic of the most sumptuous Jacobean interiors, the Long Gallery at Aston Hall, Birmingham, 1618. In the Birmingham Museum and Art Gallery, England.

By courtesy of the Birmingham Museum and Art Gallery, England

is Italian, but the final result—with wide oak-boarded floor, and white- and gold-plastered and panelled walls designed to accommodate portraits by Van Dyke, the white marble fireplace, and the Corinthian doorcases—is truly English. The coved and painted ceiling, executed by Edward Pierce and Emanuel de Critz, plays a vital part in balancing the proportions of the room. Though Renaissance principles are demonstrated in design such as this, they were not fully developed in the country at large until the 18th century and the advent of the Palladian school of architecture and decoration (influenced by the 16th-century Italian architect Andrea Palladio).

A.F. Kersting



Figure 36: Double Cube Room at Wilton, Wiltshire, designed by Inigo Jones, *c.* 1649.

After the unsettled period of the Commonwealth, the Restoration introduced new Baroque influences from the Continent. These were fused with the restraining classicism (which was still considered to be a new style) to produce a successful balance of contrast. The designs of the great architect Sir Christopher Wren, though mainly for church and monumental buildings, relied for a great deal of their embellishment on the work of the fine artist-craftsmen such as Grinling Gibbons, sculptor and wood-carver, and Jean Tijou, ironworker, whose work can be seen in close association in St. Paul's Cathedral. In the many country houses, large plain-surfaced oak wall panels provided the perfect foil to the grace and liveliness of Gibbons' carved limewood swags (festoons), garlands, and picture borders, which incorporated flowers, fruit, musical instruments, cherubs, and monograms. In the words of the 18th-century writer Horace Walpole, Gibbons "gave to wood the loose and airy lightness of flowers, and chained together the various productions of the elements, with the free disorder natural to each species." At Petworth house, Sussex, Gibbons' genius may best be seen in the series of perfectly executed picture borders, which date from about 1690. Chimney pieces and doorcases were also decorated in Gibbons' manner, and similar floral motifs can be seen on the plaster ceilings at Ham house, Wiltshire. This house, relatively modest in size, represents without ostentation or extravagance the height of luxurious interior decoration in the late 17th century and incorporates many of the decorative innovations of that time. Among these are the practice of painting wood panelling in imitation of marble or wood graining and of gilding the moldings. Wall hangings include tapestry, gilt and painted leather, and silk damask; there is elaborate parquetry (floors inlaid with woods in contrasting colours).

Decorative genius of Grinling Gibbons

Paintings of allegorical subjects by Sir James Thornhill and Antonio Verrio ornament some of the more important buildings of the age, including the Painted Hall at the Royal Hospital in Greenwich, Wren's additions to Hampton Court, and the great chamber at Chatsworth House, Derbyshire. The intricate work of Daniel Marot, a French Huguenot architect who had worked for William III in Holland (see above *Northern Europe*), had a modest influence on the design of many small fittings and shelved

cabinets to display china—the collecting of which was a favourite pastime of Queen Mary II. Imported lacquer panels were sometimes used for the panelling of rooms, in accordance with the Chinese taste of the period.

In the last years of the 17th century and in the early 18th century the woodworker found his domain contracting. Through the influence of the grand tour and under the patronage of Lord Burlington, Italian influence predominated, the work of Inigo Jones was studied, and stone and stucco became more widely used, particularly in larger houses. The influence of the architect spread from the outside of the house to the interior decoration and even to the design of the furniture itself. Where wooden panelling was used, it was set in a simple framework. Pine largely replaced oak, and it was painted green, blue, brown, and other colours; walnut and mahogany were occasionally used for panelling. The increased use of stone and marble began with Sir John Vanbrugh, playwright turned architect, who, in his first commission at Castle Howard, Yorkshire (1699), showed an individual and masterly interpretation of Baroque, sculptural and yet with a certain grim epic quality. Applied decoration was kept to a minimum, a practice that he followed later at Blenheim Palace, Oxfordshire, where the severe and spacious entrance hall, with marble-paved floor, ashlar-faced (*i.e.*, faced with thin slabs of hewn stone) walls and columns, wrought-iron gallery railing, and frescoed dome, is the most impressive apartment in the building.

Stone staircases with wrought-iron balustrading came into common use, and by the latter part of the 18th century had almost entirely replaced the earlier, heavier timber stairs such as those at Wolseley Hall, Staffordshire, or Eltham Lodge, Kent, which had carved openwork balustrades or heavy timber balusters. In the smaller houses of the early 18th century, woodwork continued to provide the main decorative features. Wall panelling, moldings, window shutters, and many chimney pieces in simple painted pine echoed the comfortable elegance of the tall sash windows and well-proportioned rooms. Wealthier classes still employed Italian craftsmen, particularly for stuccowork, and the now familiar repertory of garlands, masks, and putti (cupids) was applied not only to the designs of Nicholas Hawksmoor, James Gibbs, and other architects of the quasi-Baroque group but also to the interiors of William Kent and the Palladian architects, whose influence became dominant toward the middle of the century (see Figure 13). In such houses as Holkham Hall, Norfolk, designed in strictly classical manner by Kent in 1734, can be seen the results of extensive travel by both architect and owner. The magnificent entrance hall is again one of the most important rooms, designed on the general lines of a Roman basilica with apse (recess) and side colonnades. At Houghton hall, also in Norfolk, Kent designed fine suites of furniture for Colin Campbell's interiors; these pieces are usually gilt, with acanthus scrolls, consoles, heads, and sphinxes; with feet and legs scrolled or of ball and claw type; and with upholstery in velvet or silk. The plaster ceilings are by Italian craftsmen, with gilded and painted ornament; the walls are dressed with classical plinth, pilasters, and frieze: and pedimented marble chimneypieces contain bas-relief panels above the mantelshelf.

Continuing use of wood in the 18th century

Wall hangings were of tapestry, cut velvet, or watered silk and damask. Elsewhere, hand-coloured, wood-block-printed papers and papers with flocking (pulverized cloth) were coming into use as an economical substitute.

Although the Rococo style never fully established itself in England, many interiors were influenced by the asymmetrical motifs (rocaille) found in the designs of such French decorators as Nicholas Pineau and J.A. Meissonier. The stucco and carved decoration became lighter, more fanciful, and more tortuous in design. Though many Baroque motifs were still used, they were more delicately modelled, and the Rococo style was characterized by elaborate patterns of interlacing C scrolls combined with such naturalistic ornaments as flowers, foliage, shells, and rocks, arranged subtly in asymmetrical yet balanced patterns. The plasterwork and carved panelling were often painted in light colours and the detail picked out in gold.

Closely allied to the introduction of the French rocaille was the revival of the Chinese taste, or chinoiserie, for architects and designers, in search of further novelty, turned again to China for inspiration. Books on travel and topography, notably Jean-Baptiste du Halde's *General History of China*, published in Paris in 1735 and translated into English in 1736, gave added stimulus. Pagodas, mandarin figures, icicles and dripping water, and exotic foliage and birds reached the height of Rococo invention. Chinoiserie was particularly popular for bedrooms, where elaborate chimneypieces and doorcases were set against the background of imported or imitation Chinese wallpapers, and the beds and windows were hung with Eastern textiles. Window hangings, with carved and gilded pelmets (valances), were becoming increasingly important, and at Harwood House, Yorkshire, the furniture designer Thomas Chippendale executed a series of pelmets with mock draperies also carved in wood and coloured to deceive the eye completely.

Popularity of chinoiserie

The Gothick taste, a further variation of the Rococo, was peculiar to England at this time. The Gothic Revival, engendered by antiquarian scholarship at the turn of the 17th century, later spread to literature and during the 1740s appeared in the more concrete forms of architecture and interior decoration. By the middle of the century the fashion was widely popular, and many houses, large and small, were in part Gothicized, both inside and out. As with chinoiserie, the products of this 18th-century vogue bore little resemblance to the original medieval models. Gothic details, originally worked in stone, were borrowed, adapted, often mingled with rocaille and Chinese motifs, and were executed in wood and plaster. At Strawberry Hill, Twickenham, Middlesex, Horace Walpole, leader of the "true Goths," borrowed the designs of medieval tombs and turned them to designs for fireplaces and bookcases. Though this vogue fell out of general fashion in the 1760s, a few enthusiasts remained who carried the Gothick taste through until it was vigorously revived again in the 19th century.

Gothick taste

About 1760 the Rococo style, with all its vagaries of taste, began to give way before the Neoclassical style, largely inspired and introduced by the architect Robert Adam, whose work reflected the newly awakened interest in classical remains. Adam returned from Italy in 1758, and, strongly influenced by both Roman architecture and interior decoration, he evolved a new style based on classical precedent, using as ornament a medley of paterae (plate-shaped motifs), husk chains, the ram's head, the formalized honeysuckle, and other elements. His style of interior decoration was deeply influenced by the gay and delicate patterns of arabesques and grotesque ornament that he had seen in various classical remains in Rome and that had already been copied during the Renaissance by Raphael and others. Adam strongly criticized the Burlington (Palladian) school for using heavy architectural features in their interiors and replaced them with delicate ornament in plaster, wood, marble, and painting, against which, in its turn, criticism was levelled. Much of his work, it may be said, is applied decoration—pretty but without basic architectural quality. With Adam, the despotism of the architect over the craftsman was complete. No detail of decoration or furnishing escaped him; his rapid and precise draftsmanship covered the whole scheme, from the overall treatment of the walls and ceiling to the decorative details of the pelmets and grates. Even carpets were made to order, and often they repeated or echoed the design of the ceiling, bringing the whole room into harmony, as in the green drawing room in the manor house of Osterley Park in Middlesex or in the dining room at Saltram House in Devonshire (Figure 37). Wood was not often left unpainted, and, although the joinery was still admirable, the enrichment was frequently in composition or metal inlay. There were especially designed templefronted bookcases, and the plasterwork was often made a frame for the decorative paintings of such artists as Antonio Zucchi or Angelica Kauffmann.

Contributions of Robert Adam

At this time, cheaper and quicker methods of decoration began to be introduced; a considerable amount of the plaster decoration was cast from molds, and a composite imitation marble called scagliola was sometimes used for



Figure 37: Neoclassical early style dining room at Saltram House, Devon, designed by Robert Adam, plasterwork and paintings by Antonio Zucchi, 1768.
A. F. Kersting

floors and columns, while cheaper woods were disguised by marbling and graining.

At the close of the century the Neoclassical style was further refined, the plaster relief decoration being simplified and lightened. The best of this style, strongly influenced by French decoration, can be seen in the work of the architect Henry Holland, who enlarged Carlton House, London, for the Prince Regent and built Southill in Bedfordshire. Holland, like Adam, was inspired by the classical monuments in Italy, where for some time he maintained a draftsman whose drawings of classical detail Holland incorporated in his plasterwork.

United States. The story of the domestic interior and its decoration in the United States is inseparable both from its own architectural development and from the story of English architecture and decoration, from which it was largely derived even long after the American Revolution. Any discussion of United States decorative design, therefore, must refer constantly to the architectural ideas that prompted change on both sides of the Atlantic Ocean.

Contrary to popular legend, the log cabin was not the earliest shelter of the first English settlers. The turfed-over dugout hut of mud-chinked saplings, not unlike the Indian wigwam with the addition of a clay-daubed wooden chimney at one end, was probably the first home of the settlers in both Jamestown and Plymouth.

Early domestic interiors

These primitive dwellings were speedily replaced by frame structures, copying the traditional small house of southeast England. At first a single room was flanked by a massive chimney (where brick quickly replaced wood and clay), but a second room was soon added on the opposite side of the chimney. The attic, later expanded into an overhanging second story, was reached by narrow winding stairs between the central entranceway and the chimney stack.

This development in New England is well represented by such vestiges as the Capen House, Topsfield, Massachusetts (1683) or the Old Iron Works (ironmaster's) House, Saugus, Massachusetts (1636). The interior clearly reflects the structure, with its massive exposed oak corner posts, beams, and joists and its huge open fireplace, which served as the cooking and heating centre of the household. Inside walls were usually of undecorated lath and plaster, covering the studs and their clay or brick filling. Windows were small and originally of casement type, with small leaded panes in a wood frame. Small windows with low ceilings conserved heat in the severe winters. Floors of wide riven boards of pine, smoothed and sanded, replaced the beaten clay of the first shelters (Figure 38).

The furniture, with few exceptions, was simple and sparse. It was decorated with simple carved and turned ornament and touches of earth colours.

By the end of the 17th century, homespun textiles were supplemented by imported woven materials in the houses of the more affluent; these were used for curtains, table covers, bed hangings, and seat pads. Richly coloured damasks and velvets, enhanced by the unpainted wood and plaster surfaces, were found in Puritan New England and, probably to a greater extent, among the less austere New York Dutch and the comparatively wealthy tobacco planters of Virginia.

In houses south of New England, brick and stone tended to replace wood as a building material, though there were many smaller timber structures that have largely disappeared. In the Hudson River region, the traditional cottage of the Flemish and Huguenot settlers, long and low with steep pitched roof and extended eaves, became the typical farmhouse. At the same time, the narrow Dutch town house of brick with its stepped gable ends gave New Amsterdam, even after the English occupation, an appearance completely different from that of the English settlements to the north and south.

In the Dutch houses, windows tended to be larger and ceilings higher. The early fireplace, with its tiled border, surmounted with a deep hood, was flush with the wall instead of deeply recessed. Dutch features such as the horizontally divided door, the monumental cupboard, or *kas*, the built-in bed, and tiling and dishes of delftware gave the early New York interior an individuality that withstood English influence until well into the following century.

Dutch features in New York

Similar national characteristics must have distinguished the early Swedish settlement on the Delaware, where, later in the century, the log cabin of the pioneer may have first appeared. The Swedish contribution was only temporary, for the settlement was absorbed by both the Dutch and the English. The early settlements of the English in east New Jersey were founded by migrants from New England who at first designed typical central-chimney houses but before the end of the century largely abandoned them for the Flemish type of house in the neighbouring Hudson River region. The first settlers in Pennsylvania, arriving in Philadelphia at the end of the century, built the type of town dwelling devised for the rebuilding of London after the Great Fire of 1666.

In Virginia and the South, scant evidence remains of the early 17th-century house. Bacon's Castle in Surry County, Virginia, with its projecting two-story porch in front and

Richard Merrill



Figure 38: Simply furnished New England domestic interior: Great Room, Old Iron Works (ironmaster's) House, Saugus, Massachusetts, 1636.

rear stair tower, built in brick about 1665, is all that remains of a colonial version of the small English Jacobean manor, though there must have been several other examples. From surviving evidence and deduction it is believed that panelled walls, carefully designed beamed ceilings, and ornamental plasterwork in colour were employed in larger Virginia houses. Yet, while the milder climate made loftier ceilings and larger rooms possible, it is unlikely that the ordinary early dwelling differed from its Northern contemporary except in its greater use of brick and in placing chimneys at the ends instead of at the centre of the structure.

Among the wealthy the principal articles of furniture were undoubtedly English imports; the more humble settler probably had to make do with articles of the simplest sort, but since few articles survive from this period, little is known about it. Certainly the scattered or rural character of the Southern settlements and their concentration on tobacco planting failed to encourage the early development of skilled crafts found in villages and towns of the Northern communities. By 1720 the design innovations of Inigo Jones and Sir Christopher Wren, as reflected in the Queen Anne style with its strong mingling of Dutch and Flemish elements, had already crossed the Atlantic. Wren's influence is increasingly evident in the tendency to employ symmetrical design around an accented central feature and, particularly in the interiors, in the greater insistence on classic arrangement in the positions of openings and of panelling. Panelling, usually of pine in the north, was generally painted. Relatively deep and strong tones—red, blue, green, brown, and yellow—were used either singly or in combination, producing an effective background for the walnut furniture of the period.

Influence of Wren and Jones

Additional colour was introduced by more elaborate use of woven and embroidered textiles, in upholstery as well as draperies. Though woven carpets for floor coverings were rare even at midcentury, frequently their effect was achieved by stretched canvas painted with allover repeat patterns.

Throughout the colonies, furniture became more plentiful and varied. Chairs without arms took the place of stools, the cabriole (curved leg) largely replaced the turned leg, and small drop-leaf tables replaced the fixed-frame type. Bedroom furniture became differentiated with the development of the high chest (highboy) and the dressing table (lowboy), and later the case-top desk or secretary became the principal ornament of the living room. Tall mirrors with crested tops replaced the small, square, Jacobean style looking glasses of the 17th century, and portraits and prints came into more general use, sharing the wall space with bracketed candle holders or sconces. Artificial light still came mainly from small wick and grease lamps, but tallow and wax candles held in sconces, in adjustable metal and wood floor stands, or in candlesticks of brass or pewter (and occasionally in brass chandeliers) were used by the wealthier.

Colonial furniture

Though domestic comfort was improving, north of Virginia the large formal house or mansion remained a rarity until about 1750. In the South the wealth of the slave-holding planter made it possible for him to copy the early Georgian type of manor house in England. Great houses of two or three stories with side dependencies (outbuildings) became numerous. Stratford in Westmoreland County and Westover in Charles City County, Virginia, built about 1735 by the Lee and Byrd families, are early examples of the type. The elaborately panelled rooms of these mansions were furnished according to the latest London fashion. Probably only later in the century were these English pieces mingled with those from the cabinetmakers of Philadelphia, New York, and Boston. Between 1750 and the Revolution this Georgian phase reached its highest development. Though generally smaller and lacking the forecourt and dependencies of the southern mansion, the larger houses of the north, such as the Wentworth house in Portsmouth, New Hampshire, mark perhaps the most distinctive achievements of colonial design and decoration by their apt translations into wood of brick and stone Georgian forms.

In the Middle Atlantic colonies, particularly in Philadelphia (which by 1760 had assumed urban leadership in the colonies), a type of domestic design midway between that of New England and Virginia had developed. There the English Rococo decorative style publicized by Thomas Chippendale received its most competent and original interpretation. This is well seen in Philadelphia interiors such as those of the Powel House (1765) and Mount Pleasant (1762) and in the work of cabinetmakers such as Thomas Affleck and Benjamin Randolph (Figure 39). By this time mahogany, with its fine grain, so receptive to carving and high finish, had largely replaced walnut as the principal cabinet wood. Inspired by this material and the challenge of London design, these Philadelphia craftsmen and their northern contemporaries, particularly John Goddard and Job Townsend of Newport, Rhode Island, brought their art to the highest level of perfection.

English Rococo in Philadelphia

By courtesy of the Philadelphia Museum of Art



Figure 39: Middle Atlantic adaptation of the English Rococo style using Philadelphia Chippendale furniture: Great Chamber at Mount Pleasant, Philadelphia, 1762

During the third quarter of the 18th century, the panelled interior reached its most elaborate form in the colonies. North of Virginia a fully panelled room was exceptional; wood panelling was reserved for the chimney breast and its flanking recesses or cupboards. In Virginia and the South, full panelling remained the rule. (At colonial Williamsburg, Virginia, surviving houses have been carefully restored and furnished, giving a complete picture of the comfortable panelled rooms dating from the middle decades of the 18th century.) In both North and South, however, the mantel and its overmantel were emphasized as a decorative unit, and the Baroque broken pediment became the usual crowning feature of both overmantel and doorway. Painted woodwork remained popular, but with softer and lighter tones, tending toward white and gray. Plaster wall surfaces were also painted. Block-printed and painted wallpapers were frequently used in the main rooms of these houses, and there are indications that fabric wall hangings were used also.

Plaster ceilings completely concealed the floor beams by the second quarter of the century, and after 1750 these were frequently decorated with ornament in low relief in the French or Rococo manner and hung with many-branched chandeliers of crystal. Floors of hardwood, occasionally parquetry, were more frequently covered with patterned rugs of European or Oriental origin.

During the 18th century imports of printed cottons or chintz in the Indian taste, and silk brocades and damasks, largely replaced the linen and woolen weaves of earlier days. Upholstered furniture, wing chairs and sofas, and elaborate draperies increased still further the richness of the fashionable interior.

Use of textiles

As in Europe, the growth of tea and coffee drinking encouraged production of suitable silverware and the import of English and Oriental porcelains, which required corner

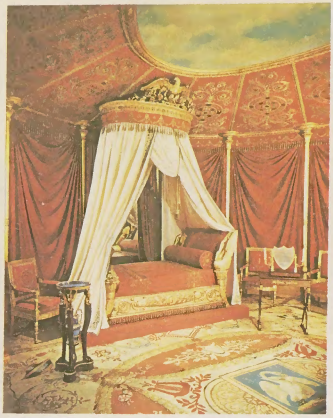


Figure 40: Roman decorative motifs characteristic of the Empire style: bedroom of the empress Josephine in the Château de Malmaison, near Paris, 1810.

By courtesy of the Musee National du Chateau de Malmaison; photograph, Studio Laverton

and wall storage cupboards. Demand was also created for a variety of small movable tables and stands for tea and coffee services.

During this century the German settlers in Pennsylvania added their traditional styles of design to the dominantly English tradition of the colony, the effects being more evident in folk arts than in formal decoration. It was to this style and its development after the Revolution that the first American decorative glass of Henry William Stiegel and Frederick Amelung must be credited, as well as most of the decoration on early American pottery.

19th and early 20th centuries in Europe. Neoclassicism predominated in France till the rise of Napoleon, when to Roman styles were added Egyptian motifs from his Egyptian campaign of 1798. This was known in France as the Empire style, after the First Empire of France (1804–14), and in England as Regency, for the period (1811–20) when George III was too deranged to rule. Furniture design, for the most part light and graceful during the early part of the Neoclassical period in France, had become more consciously luxurious as the Revolution was approached. During the Empire period it became massive, imposing, dark, and pompous (Figure 40). The usual vocabulary of classical ornament is to be found in both Empire and Regency, with some modifications from earlier times. The cabriole leg of the Rococo style became straight, and curves tended to disappear in all furniture. Symmetry of ornament replaced the asymmetrical curves. In England, in the latter part of the 18th century, porcelain became less and less fashionable, and its place was taken by the cream-coloured earthenware (creamware) of Josiah Wedgwood, and by his jasper and basaltes stonewares, all admirably adapted to the new style. Greek vase-shapes and classical ornament were commonly used in the decoration of Wedgwood wares of all kinds. In England, the work of Thomas Hope, a wealthy amateur architect, gained much attention through the publication of his *Household Furniture and Interior Decoration* (1807). He enlarged and decorated his London home in Duchess Street, Portland Place, and also his country house, Deepdene, in Dorking, Surrey, with somewhat heavy and pedantic design that was at variance with the general trend of the time but influenced later work.

In Germany the solid bulk of the Biedermeier style, with its thick curtains, draperies, antimacassars, and padded upholstery, gave evidence of material prosperity. Many of these features were to become commonplace in Victorian England, but in the meantime, the Regency style was prevalent and contributed many masterpieces of design. Brighton Pavilion (begun 1815) was built by John Nash for the Prince Regent. Much lacquered and bamboo furniture was used, blending with Chinese wallpapers, fanciful treatments of palm trees as columns, and the most extravagant of crystal chandeliers (Figure 41). In general, however, the Regency style strove for elegance without extravagance; innumerable smaller houses were built and decorated with fine wrought-iron balustrades on curving stone staircases, pleasing carved wood or marble mantelpieces of modest

The Regency style in England

By courtesy of the Royal Pavilion, Brighton, England



Figure 41: Regency style interior using bamboo and lacquered furniture and decorated with chinoiserie motifs: the Prince Regent's bedroom, Royal Brighton Pavilion, Brighton, England, designed by John Nash, begun 1815.

sizes, and plain or panelled walls of light colouring, on which the use of wallpaper was becoming more common.

By the latter part of the 18th century, the Industrial Revolution was slowly developing, particularly in England, and machinery was increasingly producing many objects of interior decoration, modifying their form to suit the new methods and reducing the price to make them available to new markets, a situation envisaged by Wedgwood. The less affluent of the middle classes became the largest section of consumers, and manufacture was increasingly directed toward catering to their tastes. In the early years of the 19th century a new concept was beginning to take shape—the notion of eclecticism, which propounded that any style was as good as another. This led to the idea that styles could legitimately be mixed together. In this way Horace Walpole's nightmare of a garden-seat—Gothic at one end and Chinese at the other—became, in principle, an accomplished fact: one firm, for instance, made a classical urn on a Gothic base.

Eclecticism as a style

In the early decades of the 19th century, in addition to the Empire and Regency styles, there was a Greek style of marked simplicity, and an Italian style described as 'picturesque with Palladian detail' (a contradiction in terms), as well as an "Elizabethan" style, a "Tudor" style, a "Baronial" style (under the influence of Sir Walter Scott), an "Abbotsford" style (also resulting from Scott's influence, based on his house of that name), and a revived Gothic style, far removed from Walpole's modest and amusing essay. The revived Gothic was at first inspired by James Wyatt's pseudo-cathedral built for the author William Beckford at Fonthill Abbey, with interiors of cathedral-like amplitude and about a 300-foot (90-metre) tower.

The Gothic Revival

This Gothic Revival produced a small number of houses in which the pointed arch together with fan vaulting and crocketed (carved with foliated ornament) or deeply undercut mouldings were used with some taste and discretion. Toddington Manor, Gloucestershire (1829), by the architect Charles Barry (who, with A.W.N. Pugin, designed the Houses of Parliament), and Highenden Manor, the house of British prime minister Benjamin Disraeli, exemplify a style used later in the century with greater ostentation and coarseness of detail.

In the principal European countries, interior decoration grew increasingly heavy and elaborate. Ornament came to be considered synonymous with beauty, and pattern covered every possible surface. The products of industrial manufacture were mostly very crude, and their use resulted in loss of refinement; for example, aniline dyes, which are harsh in colour, were first made in 1856 and soon replaced the softer, more harmonious colours. Architects decked out their buildings according to whim in a variety of styles.

In less ambitious schemes of decoration brightly coloured wallpapers with bold patterns were widely used, and the white plaster ceilings were relieved by modelled cornices and often also by some central feature, frequently in a coarsened Rococo design, which made a background for the elaborate light fitting. Rooms became crowded with furniture, and fireplaces were often mounted with elaborate overmantels, fitted with mirror panels and a multitude of shelves and brackets for the display of knick-knacks. Both furniture and fittings were draped in dark-coloured plush with heavy fringes. Varnished pitch-pine dadoes, stained-glass windows, and encaustic-tiled floors were also popular.

By the 1830s there was a revival of Rococo, to be seen in the porcelain of the period and the chairs of John Belter of New York, and there was something called the "Louis XIV" style, which that monarch would have found difficulty in recognizing. Throughout this period there was a limited amount of pseudo-Chinese decoration, principally on pottery and porcelain and papier-mâché. After 1853, when Commodore Matthew C. Perry of the U.S. Navy reopened Japan to Western trade and influence, a new kind of Japanese art began to be exported, such as the vases of unprecedented ugliness decorated in Tokyo and called Satsuma, or enormous, grossly over-decorated vases from Seto in Owari (presently Aichi Prefecture), none of which would have found a buyer in the Japanese home-market.

Oriental Rococo

The 19th century was an age of eclecticism. Decorators introduced the custom of having a different style for each room—"Gothic," "Elizabethan," or "Old English" for the dining-room; "Queen Anne," "Chippendale," or "Louis XVI" for the drawing-room; with pseudo-Elizabethan furniture for the library. Design reached its nadir with the Great Exhibition of 1851, in London, the low-water mark in the history of European taste in interior decoration, from which there was no conceivable direction except upward.

In France, where there was a sounder tradition and Gothic had not been influential for centuries, 19th century taste was not quite so debased as in England. A light and amusing version of Gothic known as the Troubadour style made its appearance in the 1830s, perhaps an international tribute to the contemporary fame of Sir Walter Scott. Rococo was revived as the Pompadour style, and there was a neo-Renaissance period, with furniture designs based on 16th-century Italian work. On the whole, the furniture of the second empire (1852–70) was very acceptable in design, although these pieces were based largely on the 18th century; these styles harmonized well with the contemporaneous music of Jacques Offenbach and the brilliance of the court of Napoleon III.

In England there were a few people who recognized the depths to which taste had fallen. The designer and writer William Morris advocated a return to fine craftsmanship in furniture, textiles, and wallpaper, and started his own firm in 1861. Under the influence of the Pre-Raphaelite Brotherhood, artists who advocated a return to medieval principles, his furniture designs were based on actual surviving specimens instead of on Gothic architecture of the most florid periods. Morris's productions were well-made and well-proportioned, often with painted decoration in the old style (Figure 42). He helped to organize the Arts and Crafts Society with the object of improving design. His influence was limited, however, because, like his contemporaries, he looked backward for inspiration and in doing so refused to accept the possibilities of machine production.

Influence of William Morris

The 1870s and 1880s saw a fashion for reproductions of 18th-century furniture, especially the designs of Chippendale, Hepplewhite, and Sheraton, in which a few minor crudities, of a kind thought to be inseparable from handwork, were added to machine-production. Much of the "18th century" furniture that decorates today's interiors is no older than this vogue. A fashion arose in the 1880s for Japanese fans and screens and blue and white porcelain, in conjunction with bamboo and lacquer furniture, a taste to some extent influenced by the paintings of James Whistler.

By courtesy of the Victoria and Albert Museum, London: photograph, John Webb



Figure 42: Outstanding craftsmanship and design based upon medieval aesthetic principles: mid-19th century arts and crafts movement English room decorated by William Morris with furniture by Philip Webb. In the Victoria and Albert Museum, London.

The influence of Whistler, Morris, and others may be seen in the Art Nouveau style of decoration, which was developed in the 1890s by the Belgian architect and designer Henry van de Velde and the British designer Arthur Heygate Mackmurdo. This was a style in interior decoration which went under various names at the time—Art Nouveau in England, Modern Style in France, the Jugendstil in Germany, and the Stile Liberty in Italy, in reference to the influence of the London firm of Liberty & Co. in promoting the style. Art Nouveau was most reminiscent of Gothic, with overtones of the Japanese art imported during the last quarter of the 19th century. Its ornament is markedly asymmetrical, and principally floral, particular use being made of the lily. It is strongly curvilinear, and there is hardly a straight line to be seen. It often derives its effect from an incongruous juxtaposition of decorative motifs. In furniture, for instance, the asymmetry of Rococo is to be found in its ornament, but in Art Nouveau the whole piece of furniture in some cases is asymmetrical, one side being higher than the other. Although the style created much interest at the Paris Exhibition of 1900, it never became very widely established but was one of several leavening agents in the sphere of design. Nonetheless, its influence extended beyond World War I into the 1920s, when the Art Deco style from Paris became current (see below). Its influence can also be found in such relatively modern designs as the Barcelona chair of Mies van der Rohe of 1929.

Reaction against overcrowded, fussy interiors gathered strength. Plain interior walls in white or very light colours, natural woods, and simple doors and fireplaces were among the changes introduced by the more advanced designers in an attempt to create an original style suited to the changed circumstances of life in the first part of the 20th century.

Late 18th to early 20th centuries in the U.S. *Classic movement after the Revolution, 1785–1835.* Even after the American Revolution, English decorative influence predominated in the United States, in spite of greatly increased contacts with French thought and ideas. Although many leaders like Thomas Jefferson wished to see a complete break with English traditions, the Georgian forms of colonial days persisted in common usage till 1800 or after. By 1785, however, the reaction in Europe against the rather heavy classic style called free Palladianism and its Rococo and Baroque elaborations began to affect design in the United States.

Jefferson, largely under French influence, became the leader of one aspect of the new movement in the South that combined practical planning with a literal classicism based on the direct study of ancient monuments. While Jefferson's interest in strict classic form was felt particularly in architecture, the decorative phase of the movement, both North and South, was dominated by the freer and more personal interpretation of classic motifs based on the work of the Adam brothers in England, before and during the American Revolution. This was the principal influence in the designs of the Boston architect Charles Bulfinch and his followers and was popularized about 1800 in the builders' pattern books of William Pain and Asher Benjamin.

The houses of Boston, Salem, and Portsmouth that were built around 1800–10 by or under the influence of Bulfinch and Samuel McIntire, an architect of Salem, are the best examples of the changes wrought by the fine scale and delicate precision of their Adam-inspired designs, producing what has become known as the early Federal style. In the houses of the time, the circle, the ellipse, and the octagon were introduced as occasional variations in the plan, and the flying or freestanding staircase became a characteristic of the entrance hall (Figure 43). Federal style

In interior decoration, wood panelling was practically abandoned or was restricted to the area below the chair rail—*i.e.,* the wall molding at the height of the chair back. Decorative emphasis was concentrated on the mantel and overmantel, the doors and window frames, and the cornice, all usually of wood and enriched with delicate repeat ornament (either carved or applied). Rich colour in draperies and upholstery was set off by wall surfaces and decoration in light tones, grayed tints, or white. Block-printed wallpapers with classical motifs were frequently used, as were stencilled decorations in the simpler homes.

In general, geometric forms and the urn, swag, patera, and wreath were employed. The taste for lightness and attenuation verging on dryness was reflected in the furniture. The designs of the English furniture manufacturers George Hepplewhite and Thomas Sheraton, influenced by Louis XVI and Directoire forms, found American versions around the turn of the century in the work of Samuel McIntire of Salem, John Seymour of Boston, Duncan Phyfe of New York, Henry Connelly of Philadelphia, and the cabinet shops of Baltimore and Charleston. At first, light woods and finishes and decorative inlays were preferred, but by 1820 French Empire influence substituted dark reddish mahogany, carved and gilded ornament, and heavy, often ill-proportioned forms considered more in keeping with classic taste.

After 1820 the early Federal style waned, and Jeffersonian classicism was modified by the introduction of Greek and even Egyptian detail, constituting the so-called Greek Revival style

By courtesy of Antiques Magazine; photograph, Helga Studio



Figure 43: American neoclassical room in the manner of the Adam brothers: Oval Music Room, Nathaniel Russell House, Charleston, South Carolina, c. 1800.

Revival. Accompanied by furnishings and draperies in the heavier Sheraton-Empire taste, the classic pattern established in the 1820s became the basic style in building and decorative design. Stimulated by the Greek struggle for national independence, it lasted until about 1850 and constituted for the time a national style without parallel in Europe. In its later decorative aspect, however, the Greek Revival became a fashion rather than a style. As such it marks not only the end of the 18th-century Neoclassicism but the beginning of the Romantic movement.

The Romantic movement and the battle of the styles, 1835–1925. The ordered symbolism of the Roman classic style had been envisaged by Jefferson as a proper expression of the American national ideal; but by 1835 its restraints had grown tedious. Social and economic changes already initiated by the Industrial Revolution encouraged reaction. This found more or less romantic and emotional expression in a series of style revivals ill-adapted to actual conditions.

The Greek Revival was diluted almost immediately by the antiquarian Romanticism of the "Gothic," "Tuscan," and "country cottage" fashions. These offered opportunity to the undercurrent of practical utilitarianism, repressed or thwarted by the classic formula, and also gave a fertile field for the novel or exotic in decorative taste fostered by a wealth-induced appetite for comfort and display (see Figure 2, left and Figure 9). By the middle of the century the last vestiges of order in early Victorian Romanticism had disappeared under a plethora of decorative motifs and objects easily and inexpensively produced by machine (Figure 44). Colour became confusedly drab or brilliant and generally out of character, as a result of the introduction of uncontrolled chemical dyes and the magic of the Jacquard loom, which permitted the weaving of intricate patterns. Increased travel and ease of communications made American styles hardly distinguishable from those of Europe.

Machine-produced decoration

This decorative salad of classic and medieval motifs was supplanted by the revival of the 18th-century forms which temporarily triumphed in the "second Rococo" of the 1850s, when rosewood and walnut took the place of mahogany. This was succeeded by fashions based on the 17th century and the later Renaissance, until the Philadelphia Centennial Exhibition of 1876 brought to America the "craft" medievalism and a new series of more literal



Figure 44: Victorian parlour with characteristic tufted upholstered chairs, medallion portraits, corner whatnot, and floral carpeting: Robert J. Milligan House, Saratoga, New York, c. 1853. In the Brooklyn Museum.

By courtesy of the Brooklyn Museum

style revivals including that of colonial times. These in turn absorbed the exotic Eastern influence of the Aesthetic movement of the later 19th century.

In the first quarter of the 20th century this confusion culminated in antiquarianism for the wealthy and, for most people, period reproductions provided by the wholesale decorator and manufacturer. These 90 years of enormous technical and financial development are too confused and complex for further analysis here. Almost from the beginning, however, a body of criticism and rational experiment was developing both in Europe and America that was to find effective expression in the early 1920s amid the social and economic upheavals following World War I.

20th century. The principle behind a great deal of 20th century interior decoration was first expounded in Chicago in 1896 in a magazine entitled the *House Beautiful*. This journal opposed both the perpetuation of vulgar display

By courtesy of the Metropolitan Museum of Art, New York

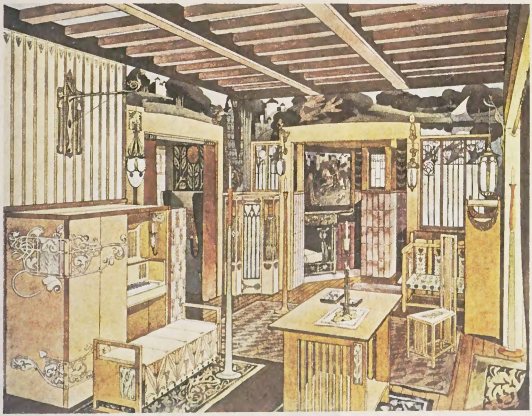


Figure 45: "Design for a Living Room," by Will Bradley, commissioned by *The Ladies Home Journal*, 1902. In the early 1900s *The Ladies Home Journal* published a series of simplified, contemporary house designs appropriate to the needs and taste of the new century. In the Metropolitan Museum of Art, New York City.

and the excess of ornament that had characterized most of the 19th century. Other American magazines like *The Ladies Home Journal* soon followed *House Beautiful*'s lead and published articles on modern decorating (Figure 45). In Europe a group of architects and designers whose thesis was that "form follows function" started the Bauhaus, a school of design founded in 1919 at Weimar, Germany. With such pioneers of modern art and design as Walter Gropius, Paul Klee, László Moholy-Nagy, and others on its staff, it sought to teach the combining of art with craft, and to combat the dehumanizing effect of the machine.

Influence of the Bauhaus

The struggle between the desire to cling to tradition and the necessity of accepting a society based on mechanized industry came into the open between World War I and World War II. The aim of the Bauhaus group was to adapt industrial techniques to meet the needs of a society impoverished spiritually and materially by war. Their work was the culmination of the numerous reform movements of the late 19th and early 20th centuries; cathartic and analytical in its methods, on one hand it shocked the conservative into immoderate fury and on the other converted its radical adherents into equally uncompromising iconoclasts. Many of the "functionalist" ideas they employed were inspired by the subtle simplicities of the Japanese tradition and by the innovations and writings of the Chicago architect Louis H. Sullivan. Functionalism demanded a complete break with the ornamental motifs of the past and a quickened response to form, proportion, line, and texture. It also aimed at a scientific study of human behaviour, correlating psychological responses to physical stimuli of all kinds. The acceptance of its thesis ran parallel to the growth of interest in abstract art, and, although the uncompromising application of so intellectual a program proved immediately impracticable, its bold challenge to convention resulted in notable changes in interior design.

The style that emerged from the Bauhaus, called the International Style, was felt by many to be lacking in human warmth. Its boxlike forms, its hard and glassy surfaces, its use of metal tubing and plywood, and its lack of colour and of ornament were received with mixed feelings. The French architect Le Corbusier adhered to similar principles. His famous dictum that the house is a machine brought the retort that most people do not like living in machines. Functionalist thinking, however, led to an increasing use of the materials the machine is capable of producing, such as plastics, synthetic fibres, acrylic paints, and so forth, but these materials were still too often used to simulate other materials.

German Functionalism was slow to establish itself in Europe and hardly affected American design until its leaders found refuge in the United States from Nazi oppression. There the movement was brought to public attention in the mid-1930s by the need for new stimuli in the trough of economic depression, by the educational campaigns of the Museum of Modern Art in New York City, and by the reestablishment of the Bauhaus teachings in the Institute of Design of the Armour Institute (now part of the Illinois Institute of Technology) in Chicago.

Art Deco style

In the decade following the Exposition Internationale des Arts Décoratifs et Industriels Modernes, held at Paris in 1925, progressive Western design was influenced principally by the less radical productions of the French luxury crafts, based on a modified Art Nouveau, and by the Swedish success in combining and developing craft traditions in cooperation with industry. These influences, which developed the Art Deco style, were, however, confined to relatively small and semiprofessional coteries, while the market as a whole continued to concentrate on traditional forms, producing and adapting them at various levels of quality and taste (Figure 46). By 1935 the Functionalist movement, led by the disciples of the Bauhaus program, had gained a substantial following among the younger architects and designers. During World War II, development virtually ceased in most European countries, and subsequently attention turned again to the Scandinavian countries, particularly Sweden, where strict consideration of function led to simple furnishing schemes which relied on natural wood grains, clear colouring, and texture for their effect. Pattern was subdued and, where used, uncomplicated in outline.

Meanwhile, in the United States, during and after World War II, the Functionalists, still with the help of the museums and the more progressive schools and periodicals, had gained the interest of a considerable proportion of both the wealthier members of society and the manufacturers who catered to them.

Effects of the Functionalist movement

The most obvious changes resulting from the Functionalist movement were mechanization, redistribution of interior space, and elimination of formal barriers between indoors and outdoors. These developments, most prevalent in the United States but disseminated throughout much of the world, were accompanied by radical changes in decoration and the design and use of furniture and fittings. Equipment for heating and lighting, sanitation, and food preparation, all derived from inventions of the 19th century, were brought to a high degree of mechanized efficiency, taking full advantage of advanced production methods. Since convenience and economy became principal considerations, utility units were fitted into living space instead of being hidden in otherwise unused areas, as in the traditional room arrangement. By insisting on simplicity of form, colour, and texture, they were made to obtrude as little as possible. In particular, the appearance of the kitchen was studied carefully, especially in smaller houses.

Under the influence of electric power, liquid fuels, flexible controls of temperature, ventilation, and lighting, and countless labour-saving devices, the mid-20th-century house began to fulfill Le Corbusier's dream of an efficient "machine for living."

Reconsideration and correlation of the space needed in living areas broke down traditional room divisions. The new interior, with its invitation to movement, both actual and implied, was in harmony with the times. Decoration became concerned with function (see Figure 2, right), and, because a living area served more than one purpose, it

Photo Fratelli Fabbri Editori, Milan, Italy



Figure 46: Art Deco bathroom designed by Armand-Albert Rateau for Jeanne Lanvin, Paris, 1920–22. In the Musée des Arts Décoratifs, Paris.

was frequently irregular in plan and impossible to treat as a unit in the traditional formal manner. Changes of colour, texture, and materials consequently became the chief resources of decorative design, taking the place of ornament (Figure 47). Earlier attempts at the functional mode suffered from too much anxiety over simplicity and unity and thus became monotonous and cold.

Photo R. Guillemot—TOP



Figure 47: Dining room and living area designed by Claude Lombardo for his apartment outside Brussels, 1969. Supple, rounded forms made of cement reinforced with glass fibre are used to create a free-moving, open-plan interior.

The demands of space made it necessary to keep movable pieces of furniture to a minimum and encouraged the use of built-in units. An earlier overemphasis on straight lines and angles was countered by greater use of curved and molded forms in furniture design. As the average house became smaller and more efficient in its use of enclosed space and as the desire for outdoor living grew, there was a tendency to replace at least one of the enclosing walls of both livingroom and bedroom with glass. With a well-arranged plan, this gave each room an everchanging mural and better light, and it also extended the apparent size of the interior. The illusion of bringing the outside indoors gave a feeling of freedom, but it also created practical and psychological problems (see Figure 7).

Despite the reaction that developed against it, the functional modern movement had served an important purpose. Although it produced no recognizable themes of ornament, it did eliminate the *horror vacui* that afflicted the Victorians and Elizabethans alike. It cleared the way for a fresh look at the art of interior decoration as a whole, and for the fresh inspiration that came in the 1950s from Scandinavia and Denmark, which retained the human qualities that much of the work of the Bauhaus was felt to lack. At the same time there was a revival of interest in true Japanese art in interior decoration, which has a certain affinity with Scandinavian. In the 1960s patterns began to return—abstract patterns such as those to be found in Op art. Elegant materials, easily washable, became available for upholstery, and easy cleaning made it practical for them to be produced in pastel shades and light colours.

That large numbers of people had found it difficult to live with modern austerity became apparent with the immense growth after World War II of the trade in old furnishings of all kinds, with ever-increasing prices. A parallel vogue resulted in an increase in the manufacture of reproductions of all kinds, especially furniture, made partly by machine and partly by hand, leading to the revival of some of the old handcrafts.

INTERIOR DESIGN IN THE EAST

East Asian motifs of decoration bear no relationship to those of the West, although many of them are familiar from *objets d'art* and decoration exported during the last five centuries. No such conflict of styles as those to be observed in the West has existed.

Eastern motifs

The motifs of Eastern art are many and varied, such as the dragon (a ubiquitous and beneficent creature), the so-called phoenix (actually the Chinese long-tailed pheasant), and creatures of all kinds, actual and legendary. Flowers and foliage are part of an elaborate flower-symbolism, and there are many abstract motifs, all of which are part of a complex and rich symbolism, which can usually be interpreted if the key is known. The Chinese language contains many identical words, which have completely different meanings that are identified in speech by intonation; the word *fu*, for example, can mean either a bat or happiness. Therefore, a decoration of bats symbolizes happiness. This is not true in the Japanese language, but the Japanese have taken over many Chinese motifs, such as the bat (*kōmori*). The purpose for which a Chinese object decorated with a dragon was originally intended may often be deduced from the number of claws to the foot—five for the Emperor, four for princes of the blood, and three for officials. The pine, willow, and bamboo in conjunction are termed the "three friends," and represent Buddha, Confucius, and Lao-tzu.

Scrolls of painting or calligraphy are characteristic of interior design in the East. They are changed from time to time to give freshness to the decorative scheme and also to emphasize their quality. Similarly, a vase with a single branch of peach blossom or other flowers may be set out with care. Cabinets and storage chests are of great importance and are often made of camphor wood. An important feature in the houses of north China and Korea is the *k'ang*, or heated brick platform, on which the family sleeps or sits in the cold northern winter.

China. Possessing the oldest Eastern civilization, China has powerfully influenced the others. Forms and motifs of decoration, which began as early as the Shang dynasty (18th to 12th century BC), or even before in the legendary Hsia dynasty, persist throughout Chinese history. Early forms of bronze altar vessels, for example, are found in porcelain in the 18th and 19th centuries, slightly altered in profile but still recognizable.

Materials are very different from those of the West. The Chinese have always been masters of the ceramic art, and their skill spread northward to Korea, northeastward to Japan, and south to the countries of Southeast Asia. Nearly all the more important techniques—majolica excepted—came from China. The T'ang dynasty (618–907) was renowned for fine earthenwares; the Sung dynasty (960–1279) for superb stonewares; and from the Yüan dynasty (1206–1368) onward the Chinese have led the world in the manufacture of porcelain, the secret of which reached Europe only after the porcelain had been imported for several centuries. Bronze was employed for vessels rather than figure sculpture. Originally purely religious in connotation, bronze vessels were given as gifts of emperors to their favoured subjects by the Chou dynasty (1111–255 BC), and from that time on were commonly employed for secular purposes. During the T'ang dynasty, handsome mirrors as well as such useful and decorative things as toilet-boxes were commonly made.

Chinese textiles

China was known for its silk in the West in ancient Roman times. Fragments of silk were found in Chinese Turkistan dating to the 1st century BC with motifs of design strongly resembling those of the 20th century. The Chinese have always been noted for superb silk embroideries, highly detailed in a manner requiring a multitude of tiny stitches. Painted silks have been produced in large quantities. Velvet weaving, usually in long strips as chair

covers, was an art probably learned from the West, but the art of tapestry (*k'o-ssu*), may go back as far as the Han dynasty (206 BC–AD 220). Carpet-knotting of the highest quality, no doubt learned from Persia, cannot be proved to date before the 17th century, but it may have started at a much earlier date. Rare carpets are knotted with silk and gold, but those with a woollen pile are of fine quality. Pillar-carpets, woven to encircle pillars, are a distinctively Chinese type. Motifs of decoration are those common to other materials.

Jade (nephrite and jadeite) is carved in China into objects with many different purposes. In early times, like bronze, it was mainly used for religious purposes, but it later came to be employed for a variety of secular objects, principally those intended to furnish the scholar's table, such as brush-pots, ink-slabs, water-droppers, table-screens, and paper-weights. In the 18th century especially, bowls and covers, handsomely carved and pierced with a variety of motifs and patterns, were made for interior decoration as incense burners.

Lacquer, the solidified sap of a tree (*Rhus vernicifera*), has been widely employed for a variety of decorative purposes on a foundation of wood or, less often, hempen fabric. Lacquer is employed as a form of paint, or applied in thick layers that can be carved with knives. It is also used to decorate structural timbers in the interior. The finest lacquer came from Japan in the 17th and 18th centuries.

Enamelling on metal is an art that the Chinese learned from Europe, but, in the 18th century especially, some very large bronze vessels in a variety of ornamental forms were covered with enamel utilizing the closionné technique. Painted enamels came from Canton in the 18th century, and resemble in style contemporary porcelain enamelling from the same place.

Paintings are usually on silk, and most are in the form of scrolls to be hung on the wall. A long and narrow form is customary. The best of Chinese painting is superb in quality, but criteria of judgment are very different from those applicable to Western art. Style is to a considerable extent affected by calligraphy, and the quality and type of brushstroke plays an essential part. Subjects are usually the poetic delineation of landscape, floral and foliate sprays, and, less often, pavilions. Chinese painting is often pervaded by a subtle and gentle humour hardly seen in Western art. Calligraphy plays an important part in the art of the East; scrolls decorated with an admired calligraphy are hung on walls. Calligraphy often plays a part in the decoration of bronzes and porcelain, and inscriptions on paintings are not uncommon.

The East Asian house is usually constructed of wood and tiles. The ridge-tile in China, made of glazed stoneware, is often very handsome. Architecture has never been the principal medium for the expression of the Chinese artistic impulse; conservatism, perhaps rooted in ancestor worship, has been paramount and stylistic innovation practically unknown. The basic structure of the Chinese house has remained almost unchanged at least from the Shang dynasty (18th to 12th century BC). In all types of buildings the roof is the most important feature, and by the T'ang dynasty (AD 618–907) the characteristic upturned eaves and heavy glazed and coloured tile covering had developed. The roof is chiefly supported by timber posts on stone or bronze bases, and the walls of the building serve merely as screens in brick or timber. Floors are often of beaten earth packed tightly into a timber border. Usually, a family house was composed of a series of buildings or pavilions enclosing a garden courtyard and surrounded by a wall. The courtyard played an immensely important part, because of the ever-present ideal that man should live in harmony with nature: a small pool with a lotus plant, a tree, and large rocks symbolized the whole natural landscape, and it was on these features that most care was lavished.

The Chinese house

The supporting pillars and brackets of important buildings were carved and painted, many of the designs being similar to those made familiar by Chinese pottery and porcelain. The yellow dragon symbolizes the power of the spirit, the tiger the forces of animal life. Windows were latticed with strips of wood in varying patterns over which translucent white paper was stretched. In addition to the lattice-work patterns, the windows themselves took on great variety of outline, for instance that of a diamond, fan, leaf, or flower. Doorways, too, were fancifully shaped in the form of the moon, lotus petal, pear, or vase, for structural support was not required from the light panel-type walls. Some walls may have been removable altogether, as they were subsequently in the Japanese house; others were of painted wood, hung with tapestries or paintings on silk and other materials.

A description of a Ming (1368–1644) home of the leisured class mentions ceilings with cloisons (compartments) in yellow reed work, papered walls and pillars, black polished flagstones, and silk hangings. Richly coloured rugs, chair covers, and cushions contrasted with dark furniture, which

By courtesy of the Philadelphia Museum of Art, given by Wright S. Ludington (in memory of his father)



Figure 48: Chinese scholar's study, Peking, late 18th or early 19th century. In the Philadelphia Museum of Art.

was arranged according to the strict ideas of asymmetrical balance.

Chinese furniture

Little is known of early Chinese furniture, apart from what may be gathered from paintings and similar sources. Low stools and tables were early in use, and chairs, dressing tables, altar tables, and canopied beds were common by the Western (early) Han dynasty (206 BC–AD 25). Designs and materials underwent very little change in the intervening years. Rosewood has always been widely employed, and in the palaces elaborate pieces were encrusted with gold and silver, jade, ivory, and mother-of-pearl. The Chinese interior was more extensively furnished with chairs, tables, couches, beds, and cabinets of cupboards and drawers than was the custom elsewhere in the East (Figure 48). As in Europe, the chair with arms was thought to be a seat of honour. The woods employed are native to the country and were hardly ever exported to the West, though Chinese rosewood is fairly well known in the West because most exported furniture was in this wood. Carved lacquer furniture, like the throne of Ch'ien Lung in the Victoria and Albert Museum, London, was reserved for the emperor and high officials, and the massive incised lacquer screens, known in the West as Coromandel screens, were occasionally exported. Furniture of bamboo, principally intended for garden use, has hardly survived, but barrel-shaped seats of porcelain for the same purpose are not uncommon. Carved decoration on furniture is nearly always extremely simple in design and limited to some form of interlacing fret.

Japan. Interior decoration in Japan was much influenced by Chinese ideas, especially between the 8th and 12th centuries, but it developed along lighter, more austere and elegant lines. It has altered little since medieval days. The most important differences in modern design are that the matting has been extended to cover the whole of the wooden floor, and sliding doors have replaced single-leaf screens or curtains. Two sides of a Japanese house frequently have no permanent walls, and interior partitions are of paper on a wood frame which admits a soft, diffused light. These partitions are usually moveable, allowing the interior to be rearranged (Figure 49).

Comparison of Chinese and Japanese design

The Japanese interior is a carefully thought-out arrangement. Wall-decoration hardly exists, and the walls provide a neutral background for the rest. Since the Japanese invariably cover their floors with rice-straw mats and sit on them instead of on chairs, tables are low, and are also used as an arm-rest. Tiers of shelves are common, usually covered with lacquer, and painted decoratively. They occur in a variety of forms, and the asymmetrical quality of Japanese art may be seen in these pieces of furniture, the number and position of the shelves differing on either side, and set at different heights.

In contrast to Western practice, the Japanese do not decorate their rooms with several works of art, but have a special place in the room, a focal point, at which one work of quality is displayed, and this is changed from time to time. Both the Chinese and the Japanese venerate the work of former times, and the Japanese possess the oldest art collection in the world, in the Shōsō-in repository at Nara, which was formed in the 9th century AD.

At that time, doors were pivoted in the Chinese manner, and instead of the sliding *shōji*, windows were made of wooden latticing that pushed outward, as may still be seen in shrines and temples. There was a curtained dais for the most important person and separate mats on the wooden floor for others. Then, as now, there was a connecting corridor outside the rooms. The Seiryo-den, or ordinary residence of the sovereign in the Kyōto Imperial Palace, belonged to this period and was reconstructed in the 19th century on the model of the original. A present-day family could live quite comfortably in its simple suite of rooms with walls and standing screens decorated with pictures in the Chinese classic manner.

Late in the 15th century the interior began to assume its present form as a result of a slow blending of the older court style with the more austere type of house favoured by the military caste, which was much influenced by Zen Buddhist architecture. Toward the end of the 16th century came the rise of the tea masters. These connoisseurs of the "way of tea," which involves the construction of the tearoom and its garden and correct deportment in them, established hereditary families and schools who remained the aesthetic advisers on most aspects of domestic architecture, interior decoration, and garden planning. They aimed to achieve beauty with frugality, asymmetry, and economy of movement, and much of the simple grace of Japanese interiors is due to them.

Influence of the tea ceremony

In a modern Japanese house built in traditional style, decoration is almost entirely structural, and the residences of all classes are equally neat and free from vulgarity. Their harmony and delicacy derive from an endless variation of detail in a setting that is completely standardized. Ordinary rooms are reckoned in terms of multiples of the floor-mat unit, six by three feet (1.8 metres by 0.9 metre); the sliding doors five feet eight inches (1.7 metres) high by three feet wide; the supporting pillars four to five inches (10 to 13 centimetres) square, set at six-foot intervals; and the ceiling boards one foot to 1.5 feet (30 to 45 centimetres) wide. All woodwork is unpainted and rarely lacquered, but there is great variety in the *fusuma*, or sliding doors, which divide the rooms and which are covered with paper of many patterns or decorated with paintings or calligraphy. Thus, the whole side of a room may present a landscape either in black and white or in

By courtesy of the Smithsonian Institution, Freer Gallery of Art, Washington, D.C.



Figure 49. Japanese pleasure house: "Moonlight Revelry at the Dozō Sagami," ink, colour, and gold on paper by Kitagawa Utamaro (18th century). In the Freer Gallery of Art, Washington, D.C.

colours, often on a silver or gold background. A change of these *fusuma* will alter completely the appearance of a room, and their removal will convert two or more rooms into one. All rooms can be used as bedrooms, since the bedding is stored in spacious cupboards. The reception rooms provide more scope for decoration than the others, for one end of the room is occupied by a *toko-no-ma*, an alcove with a canopy above it supported by a pillar of fine or uncommon wood, in which is hung the picture or set of pictures that, with the flower arrangement that usually accompanies it, is the only ornament. Both are changed frequently according to the season or mood. Next to the *toko-no-ma*, there is often a built-in writing table. Beside this is usually a *chigai-dana*, an asymmetric arrangement of cupboards and shelves somewhat like a sideboard. Between the top of the *fusuma* and the ceiling is often a *ramma*, an openwork frieze carved with patterns or landscapes in wood or bamboo. A framed tablet with a poem or painting on it sometimes may be placed there. Other walls are of plain plaster in subdued shades, mostly of gray or brown. The ceilings are usually of thin boards, slightly overlapping, upheld by bars about an inch (three centimetres) square, the whole suspended from the roof or floor beams. In large apartments, as in shrines and temples, the coved and coffered "Chinese ceiling," with lacquered woodwork and pictures and patterns in the coffers, is sometimes found. Fancy varieties made of bamboo and reeds and plaited wood are not uncommon. Bamboo has many uses in the Japanese house as pillars and window bars and ceiling material, when split and flattened, it may take the place of boards. Windows are of many shapes—round, square, bell-shaped, jar-shaped, gourd-shaped, diamond-shaped, fan-shaped, and purely asymmetric—and make centres of interest in a blank wall.

Types of ceilings

The furniture in a traditional Japanese house is sparse, perhaps consisting of a cabinet of blackwood or lacquer, a low writing table or a screen, either twofold or sixfold (the latter generally in pairs), decorated with landscapes on a gold or silver background and mounted in brocade. A single-leaf screen sometimes stands in the entrance hall. Among the well-to-do, other valuables such as scroll pictures, charcoal braziers, articles of pottery, spare *fusuma*, books, and curios are kept in a detached fireproof storehouse and produced only occasionally to ensure a constant variety in the rooms. It is a principle that rooms that are only occasionally occupied may be more showy and fanciful than ordinary living rooms, and these are most often met with in hotels and restaurants and other places of entertainment. Just as much care is taken with the interiors of the bathroom as with the other rooms, and the doors and windows and walls of these are usually of excellent workmanship.

India. Words of Indian origin such as calico, chintz, and palampore indicate the importance of Indian textiles in the history of western interior design. Yet the Indians themselves have never been very conscious of this role, their own domestic interiors being of the utmost simplicity, with hardly more than a carpet or prayer mat to offset stone floors and plain white walls (Figure 50). The impermanence of the materials used for the majority of dwellings may have been a contributory factor. In more palatial buildings, however, and commonly in both Hindu and Buddhist temples, walls were painted, a practice that, according to literary references, may go back to the Maurya period (321–185 BC). Paintings that survive in cave temples of the Gupta period (AD 320–600) usually depict groups of active mythical or human figures and are characterized by their sinuous lines. A late example occurs in the unfinished early 17th-century murals of the Mattâncheri palace, Cochin, Madras. Inlay of semiprecious stones, carved and bracketed pillars and capitals, and openwork marble panels also adorned the palaces of local rulers.

(Ge.S./Ed.)



Figure 50: Simplicity of domestic Indian interior: Chamba school miniature of a lady suffering the sorrows of love, late 8th century. In the National Museum of India, New Delhi.

Smeets Lithographers, Weert, Holland

FURNITURE AND ACCESSORY FURNISHINGS

The word furniture comes from the French *fourniture*, which means equipment. In most other European languages, however, the corresponding word (German *Möbel*; French *meuble*; Spanish *mueble*; Italian *mobile*) is derived from the Latin adjective *mobilis*, meaning movable. The continental terms describe the intrinsic character of furniture better than the English word. To be furniture, it must be movable.

In general, furniture produced in the last 5,000 years has not undergone innovative development in any profound sense. An Egyptian folding stool dating from about 1500 BC fulfills the same functional requirements and possesses the same basic features as a modern one. Only in the mid-20th century, with entirely new, synthetic materials such as plastic and completely new fabrication techniques such as casting have there been signs of a radical revision of the concept of furniture. Since furniture presupposes some degree of residential permanency, it is understandable that no independent furniture types seem to have been developed among the Africans, the Melanesians, the Eskimos in Greenland, the American Indians, or the Mongolian nomads in Asia.

This section deals with the materials, processes and techniques, ornamentation, kinds of furniture, and, finally, with the history of furniture.

General considerations

MATERIALS

Wood. Wood is the most used and possibly the best suited material for making furniture. Although there are over a hundred different kinds that can be used for furniture, some woods have natural properties that make them superior to the others.

A relatively cheap material, wood lends itself to various kinds of treatment; for example, it can be stained, painted,

Advantages of wood

gilded, and glued. It can be shaped by means of hand- or power-operated cutting and drilling tools. Heated, it can be bent to a certain extent into a predetermined shape and thereafter will retain the shape. The annual rings in wood create a structure with varying character, which in itself provides a natural ornamental surface, in which patterns can be formed by means of precalculated juxtapositions. Colours range from white, yellow, green, red, brown, grey to black through countless intermediary tones. By juxtaposing wood of different colours, extremely rich effects have been achieved, especially in the 17th and 18th centuries. Wood, if stored under favourable conditions, is durable, and pieces of furniture from the oldest civilizations—Egypt, for example—are still extant. Lastly, most wood has an aromatic scent.

By courtesy of the Metropolitan Museum of Art, New York, Rogers Fund, 1906



Figure 51: Carved wood chairs (1600s) and wood-panelled room (1682–85) from the *Schlössli* at Flims, Switzerland, and through the door an English carved oak bed (late 1500s) from Cumnor Palace, Berkshire. In the Metropolitan Museum of Art, New York City.

Developments in the sphere of craftsmanship and mechanical techniques, during the past two hundred years or so, have made furniture production both cheaper and quicker. Using timber as a basis and applying techniques such as shredding, heating and glueing, it has been possible to evolve new materials. To an increasing extent, cabinetmakers and furniture factories are using semi-manufactured wood such as veneer, carcass wood, plywood, laminated board, and hardboard (fibreboard).

Use of veneer

Veneer is a very thin layer of particularly fine wood that has been glued on to inferior wood in order to produce a smooth and beautiful surface. It would hardly be possible to achieve such a surface by using solid wood, partly because of the expense, partly because of its brittleness, and partly because the grain can never be shown off to its best advantage when the timber is cut into solid boards.

The practice of veneering furniture has been known since the time of pharaonic Egypt, but it was not fully exploited until the beginning of the 18th century. During the Rococo period, especially, great virtuosity was displayed by the craftsman in the veneering of curving, concave, and convex surfaces; for instance, as found on chests of drawers.

Veneer is made by sawing, machine-cutting, and peeling. Saw-cut veneer is best, but because of the relatively large loss of wood in the form of sawdust, it is also the most expensive. Therefore, furniture veneer, as a rule, is machine-cut.

Veneering is done on carcass wood, either in the form of a solid surface or a surface composed of several layers glued together. Old furniture is nearly always veneered on solid wood of an inferior quality to the veneer, such as beech, oak, or deal. High-quality English mahogany furniture made in the 18th century, however, was veneered with mahogany on mahogany. In the 20th century, machine-made laminated board of various thicknesses is generally used. The advantage of ready-made laminated board is that it does not shrink. Wood expands and contracts in various ways, and its strength can vary axially, radially, or tangentially; by blocking the wood—*i.e.*, glueing pieces of wood together in different directions—such differences are eliminated and equal strength is obtained both longitudinally and laterally. The characteristic feature of laminated board is that the veneer on both sides encloses a wooden board composed of narrow strips of wood glued together on edge. The board is therefore thick enough to be suitable for table tops or doors.

If laminated board consists only of single sheets of veneer glued together, it is known as plywood. Plywood is widely used in the manufacture of furniture, particularly as backing for chests and other storage pieces, for the bottoms of drawers, and for shelves.

Metal. Metals have been used since antiquity for making furniture or ornaments for furniture. Splendid Egyptian pieces, such as the thrones and stool that were found in the tomb of the youthful Tutankhamen (14th century BC), were rich in gold mounts (decorative details). In ancient Greece, bronze, iron, and silver were used for making furniture. Finds that were buried in the ashes of Pompeii and Herculaneum in Italy included tables with folding underframes and beds made partly or entirely of metal.

Throughout the Middle Ages the metal chair—for example, the 7th-century throne belonging to Dagobert I, king of the Franks—was used for special ceremonies.

Various examples of silver furniture have been preserved; not solid metal, they consist of embossed (decorated with relief) or chased (hammered) plates of silver fastened to a wooden core. Silver furniture was made for palaces in the days when monarchs amassed enormous wealth. In times of war, the silver mountings were melted down and turned into silver coins; it was thus that all the silver furniture disappeared from the royal palaces of France.

Popularity of iron furniture

During the 18th and 19th centuries, iron furniture became a typical industrial product. Iron beds in particular became popular. Because they could be easily folded up, they were much in demand as camp beds; one used by Napoleon at St. Helena is a famous example. As ordinary beds in private homes or hotels, they could be decorated with brass ornaments such as big knobs screwed onto their posts. Iron has also been used for chairs; for instance, rocking chairs or, perhaps more frequently, garden chairs that can stand out in the rain, protected only by a coat of paint.

The possibilities of steel for furniture were explored in Germany during the 1920s, notably by architects associated with the Bauhaus, where architects, designers, and artists experimented with modern materials. Experiments were made with steel springs and chromiumplated steel tubing. The genre was soon imitated, and tubular steel furniture became a symbol of functionalism. Since then, thinner tubing and plaited wire, with a resiliency similar to that found in wickerwork chairs have been used. Because of its lightness, aluminum became a furniture material.

Metal, however, is still employed primarily for locks, mounts, and hinges used on furniture or for purely ornamental purposes. In the Middle Ages, simply constructed chests demanded extensive use of iron bands to provide extra strength, and the ends of these bands were cut to form decorative shapes. Cabinets of the Renaissance and Baroque periods were decorated with mounts of pewter or bronze. Inlaid objects, decorated with material such as wood or ivory, set into the surface of the veneer furniture made at royal furniture workshops in France, especially so-called boulle furniture, were marked by an elaborate style of marquetry (patterns formed by the insertion of pieces of wood, shell, ivory, or metal into a wood veneer); they were influenced by Oriental traditions, in which blue-

tempered steel, brass, and copper were customarily used.

In the 17th and 18th centuries, especially in England and the American colonies, a refined style for furniture mounts, keyhole escutcheons (an ornamental shield around a keyhole), hinges, and the like, all based largely on Chinese models, was developed. The design of these mounts was dictated by a clear functional purpose, in contrast to contemporary French Rococo mounts, the majority of which were ornamental, often at the expense of utility. French bronze founders displayed great skill in making purely decorative mounts for the bodies of chests of drawers and protective mounts for corners and legs. No essentially new, independent forms of furniture mounts seem to have been developed since the 18th century.

Other materials. Among other secondary materials in furniture making, glass has been used in the form of mirrorglass or as a purely decorative, illusionistic element in cabinets and writing desks. Italian craftsmen have made glass furniture; that is, wooden furniture covered with silvered glass in various colours. Ivory and other forms of bone were used as inlay material in Egyptian furniture. During the 17th and 18th centuries, ivory was widely used for inlay work in cupboard doors and table tops.

Tortoiseshell was also used, as a costly inlay on a silvered ground, in furniture made during the Renaissance and Baroque periods. Mother-of-pearl has been used, particularly as inlay material and for keyhole escutcheons. Marble and, to a certain extent, plaster of paris have been used, especially in the 18th century, for the tops of chests of drawers and console tables, and in the 19th century for the tops of washstands and dressing tables.

Papier-mâché and plastics

In Victorian England, papier-mâché (a molding material made of paper pulped with glue and other additives) was used to make such items of furniture as fire screens, small tables and chairs, and clock cases. Finally, since World War II, various plastic materials have been used quite extensively in the construction of chairs with seats and backs molded in one piece and provided with a metal base.



Figure 52: Greek klismos and small footstool, marble grave Stele of Hegeso, 2nd half of the 5th century BC. In the National Archaeological Museum, Athens.

By courtesy of the National Museum, Athens, from Historia del Arte: photograph, E.D.I. Studio, Barcelona

STYLISTIC AND DECORATIVE PROCESSES AND TECHNIQUES

Constructional style and stylization. In general, furniture can be designed in two styles, one of which is constructional in that the appearance of the piece reflects the way it is put together, and the other of which is stylized in that the appearance of the piece conceals the way it is put together, the principle being to make the joints flush with adjoining members so as to give the impression that the object is made in one piece.

Examples of furniture made in a purely constructive style are forms employing wickerwork or bamboo, in which even the greatest display of imaginativeness in design and pattern serves to make the construction stronger and more resilient.

Constructional details and joints are not normally visible and are, therefore, seldom of aesthetic importance to the external appearance, but joints can be emphasized artistically. The Greek form of chair known as the klismos (Figure 52) demonstrates its joints boldly in the form of solid junctions holding the legs, seat, and stiles together. The curvature of the legs and of the backrest suggests elasticity. Extremely delicate joinery with invisible joints can be deliberately indicated by means of inlay work, examples of which can be seen in ancient Egyptian furniture.

Stick-back and tubular steel chairs are also examples of constructional styles. The stick-back chair consists of a solid seat into which the legs, back staves, and possibly the armrests are directly mortised (joined by a tenon or projecting part of one piece of wood and mortise or groove in the other piece). Furniture of bent steel tubing, particularly tables, chairs, and stools, was manufactured in Germany in the 1920s (Figure 58). In this fashion a new constructional style arose, for the steel tube, which makes smaller dimensions possible, was so strong that it opened up the possibility of completely new designs. Bent steel tubes form a resilient structure.

Stylization

In contrast to the constructional style is stylization, in which there is no internal conformity between the motifs and the strength of the joints. There have been any number of examples of stylization throughout the history of furniture. In both Egyptian and Chinese furniture the joints might be deliberately concealed by painting or lacquer. Chinese furniture can also appear stylized in the sense that it gives an impression of having been put together in a more constructive manner than is actually the case. (In other words, stylization attempts to make joints flush with adjoining members so as to give the impression of an uninterrupted, harmonious, or sensitive contour. When two pieces of wood are joined together with a modern, strong glue, the resulting joint will be so rigid that, in the event of a severe shock to the piece, the wood itself will be more likely to break than will the actual joint.)

A good example of stylization is to be found in French furniture made around the middle of the 18th century. In French Rococo commodes, only the back is straight. The serpentine front and sides meet in sharp corners, at which the joints are covered by brass mounts. The number and position of the drawers is concealed by an over-all pattern of veneer and bronze ornament that disregards the edges of the drawers. (In a number of cases the bronze mounts on the front consist of fanciful handles and keyhole escutcheons, but are never emphasized the way they are in corresponding English commodes, even in the case of false drawer fronts or drawers provided with moulding that serves to protect the veneer.) The fully developed French Rococo chair (Figure 53) with armrests has no visible joints. The back, arms, and frame form a continuous whole; the difference between supported and supporting members is concealed. There are no stretchers (horizontal rods) between the legs to strengthen the construction, which is solid enough by reason of the thick dimensions of the members that meet in the seat frame. To counteract the impression of heaviness in these essentially thick dimensions, the wood is molded to give a sensation of lightness without in any way weakening the construction. A chair of this type when painted or gilded looks as if it had been made in one piece, which is precisely the intention.

Decorative processes and techniques. Whether constructional principles are exploited as a motif or elegance of overall shape is stressed through stylization, every piece of furniture can be embellished in one way or another. A piece of furniture may be embellished by effects produced



Figure 53: French Rococo chairs by Louis Delanois (1731–92). In the Bibliothèque de l'Arsenal, Paris.
By courtesy of the Bibliotheque de l'Arsenal, Paris: photograph, Eddy van der Veen

in the structural wood itself or in another kind of wood added to the first; that is, by carving and turning or by inlay work. Alternatively, the piece can be decorated by the addition of materials other than wood, such as bronze, ivory, or marble. Finally, in the case of furniture meant for sitting or lying on, there is the possibility of textile enrichment in such forms as upholstery, loose covers, and cushions.

Carving. There are examples of furniture carving in Egypt at the time of the pyramids: animal legs of cedarwood on biers, beds, and chairs; and ducks' heads terminating the legs of folding stools. A more constructionally determined type of carving resulted in the creation of elegant headrests that took the place of pillows in a hot climate.

Whereas carving does not appear to have played a significant part in Greek and Roman furniture, it was a dominant feature of European furniture of the Middle Ages. The fronts of chests bear Gothic perpendicular tracery (decorative interlacing of lines) in imitation of the decorative stonework found in ecclesiastical architecture.

Islāmic wood carving

Another source of inspiration for carved ornaments in bourgeois furniture was the ecclesiastical wood carving found in choir stalls and altarpieces. The art of the woodcarver also flourished in Islām during the Middle Ages, especially in kiosks (open pavilions), oriel (large bay windows projecting from the wall and supported by brackets) windows, and Qurān lecterns. The most original and remarkable form of medieval carved ornamentation was the linenfold, which resembled folded sheets of linen laid on the surface of the wood. Although the motif was widely known, its origins are obscure.

During the Renaissance, wood-carvers changed motifs: new ornamental riches, partly inspired by the forms of classical antiquity, began to adorn cupboards and chests. Acanthus leaf designs, strapwork (narrow bands folded, crossed, and sometimes interlaced), Moresque designs, the auricular (resembling a flowered Alpine primrose) style, bunches of fruit, and scrollwork for over a hundred years dominated the figure-carving repertoires of European cabinetmakers.

During the 17th century the fashion for carved work at first receded but came to the fore again in the console tables (tables designed to fit against the wall), mirror frames, and high-backed chairs of Court Baroque. In striking contrast to lacquer cabinets of Japan, sumptuous, gilded carved work became popular on the stands invariably made for them in Europe.

In the 18th century, wood-carvers enjoyed a final splendid period of prosperity when the Rococo style of ornamentation called for the plastic effects obtainable through carving. Whole panels of woodwork (Figure 51), doors, mirror frames, chairs, and settees were adorned with the finest wood carving, featuring combinations of musselshell patterns and naturalistic vines and plant tendrils. Even in English furniture of more sober design there were ample opportunities for carved work; for example, in the many chairback variations in the Chippendale manner.

American cabinetmakers were particularly skillful at carving block fronts (the sides curving forward and the middle receding) on the drawers of chests of drawers, and the English at framing tea tables with piecrust (scalloped) tops.

Turned work. Turning is a process by which parts of furniture, such as legs and posts, are shaped while turning on a lathe. Turned work is found on Greco-Roman furniture. It is not certain whether the technique was actually employed in Egyptian furniture, though some members look as though they might have been turned. It was particularly in the shaping of wooden chair legs that Greek joiners used the lathe; the same sharp edges and deep molding seem to be repeated in the legs of bronze furniture. It is possibly ancient turned work traditions upheld in Byzantium that are reflected in certain chairs of medieval form found, for example, in Norway; made of pinewood, the construction consists principally of turned staves (thin bars), some with appendant loose rings, some of them fluted (grooved). Similar turned chairs were made in Wales in the 16th century. In the 17th century, turned work was concentrated on pillars for cupboards and on ball feet, but is also seen on chair and table legs, on which rich variations involving twisted and intertwining forms occur. Turned work in ivory also flourished in the 17th century. Except for the Windsor chair, or stick-back, however, the craft of the turner played no significant role in English furniture of the 18th century; it is similarly alien to French Rococo furniture.

Inlay and marquetry. Inlaid woodwork, in which decorative material such as wood or ivory is set into the surface of the veneer, has accompanied the art of furniture making for thousands of years. Ivory inlay can be seen in Egyptian furniture, particularly in small, meticulously executed toilet caskets, but it is difficult to locate in Greek and Roman furniture, today known almost exclusively from pictorial representations.

Italian intarsia work

In medieval Europe, inlay work gave way to wood carving and then experienced a rich period of development during the Renaissance in Italy. Italian intarsia (mosaic of wood) work found particular favour in panels over the backs of choir stalls and in the private studies and chapels, or oratories, of princes. An intarsia study of the Duke of Urbino, an Italian nobleman and patron of the arts, is still preserved in the palace of Urbino, and a corresponding room, originally at Gubbio, is now in the Metropolitan Museum of Art in New York. Together with illusionism, linear perspective (the technique of representing on a plane or curved surface the spatial relation of objects as they might appear to the eye), which had just been discovered, achieved triumphs in Italian intarsia work.

Ivory was used on both Renaissance and Baroque cupboards, sparingly to begin with, lavishly later on. Inlay work was especially used in the many splendid German and French cabinets of the period. In Holland and England an extremely rich form of marquetry (patterns formed by the insertion of pieces of wood, shell, ivory, or metal into the wood veneer) was developed, incorporating floral motifs in various kinds of exotic wood on walnut. English grandfather clocks made around 1700 often had richly inlaid cases. It was in France, however, during the Rococo period especially that inlay work reached unprecedented levels of quality. The serpentine sides and fronts of commodes were veneered with costly woods whose often relatively simple grain patterns formed an effective background for richly ornamented mounts of gilded bronze.

Upholstery and covers. Upholstery and covers belong to the sphere of furniture designed for sitting or lying on. From the Orient, Europeans learned the use of wickerwork, which provided a ventilated and resilient background for loose cushions. The upholstered chair is a genuinely European phenomenon that achieved its most distinguished and logical form in England during the 18th century. Poor heating systems in houses, general prosperity, and a desire for comfort were the conditions that gave rise to a number of imaginatively varied types of upholstered armchairs in which the only wood visible is in the legs, with the back

closing right up against the sitter and side wings affording protection from possible drafts.

The upholstered chair created a new effect that depended almost entirely upon the craftsmanship of the upholsterer. The upholstered chair or sofa has remained a specialty of the Anglo-Saxon world; club life in particular contributed to its popularity and resulted in heavily stuffed forms including that of the so-called chesterfield.

By mid-20th century, new materials such as foam rubber and various types of plastic composition had inspired independent methods that dispensed entirely with traditional upholstery techniques. Upholstery was succeeded by molded plastic forms and by sacks filled with plastic balls that are able to conform to the changing positions of the body.

Imagery and ornamentation. Painted and plastic images, or ornamental decoration, on furniture are secondary processes compared with construction and design. Some of the best and most expressive furniture forms, such as the Greek klismos chair and the English Windsor chair, are quite independent of imagery or ornamentation. On the other hand, no period in the history of furniture is entirely devoid of these secondary processes.

Superfluousness of furniture decoration

All furniture decoration is normally concentrated where it will not be in the way; for example, on the legs, arms, and backs of chairs; on the ends and canopies of beds; on the legs and stretchers of tables; and on all vertical surfaces of cupboards and chests of drawers. The superfluous nature of furniture decoration is particularly pronounced in forms that express rank or prestige. The thrones of kings and bishops, the seats of guild masters, beds of state, the writing desks of chief executives, and the like have all lent themselves to imagery and ornamentation; and as the functional aspect of the piece has declined, it has seemed that the amount of ornamentation has increased. Purely functional milk stools and typewriting tables are devoid of ornamentation. This division can be noted with varying clarity throughout the history of furniture.

At times the ornamentation itself has, in a sense, been functional. The decoration of the earliest examples of furniture from Mesopotamia and Egypt, for example, had a symbolic or magical function. The legs of Sumerian stools are shaped like those of an ox, which was the guardian animal of the city of Ur. Egyptian furniture shows a much wider development of furniture legs based on animal models. Three-footed stools ending in dogs' paws, folding stools with legs in the shape of ducks' heads, and bed legs in the form of lions' feet are known from a thousand years of Egyptian furniture history. Tables with lions' legs can be seen on Assyrian reliefs. Similar animal symbols are known from representations of Greek furniture. Sometimes the arms as well as the legs of Greek chairs had animal shapes—terminating, for example, in the head of a lion or a ram. It is thought likely that ceremonial seats and thrones featured animal motifs partly as a magical expression of the transference of power. This ancient tradition lived on in European furniture; for example, in thrones, where griffons, lions, and eagles played a prominent part in the decoration.

Even in the furniture of antiquity it is difficult to differentiate between the symbolic and the aesthetic in decorative features. It is clear, however, that the animal world has always been one of the primary sources of ornamental motifs in furniture. Animal legs and heads are found, for example, as terminal decorations in the French Rococo chair and imitations thereof. The animal leg played a prominent part in English furniture of the 18th century and later passed into American furniture. English cabinetmakers and chair makers devised a naturalistically carved lion's foot and a characteristic claw-and-ball foot, a motif that may stem from Chinese forms of ornamentation (not, however, on furniture) such as the dragon's claw holding a ball or a pearl. Richly carved English mahogany chairs sometimes also feature the heads of birds, lions, or dogs as terminal decorations on the arms. Although the majority of Chinese chairs and tables are supported by straight legs of rounded wood, Chinese thrones and seats for dignitaries have curved legs that, for some unknown reason, may be imitations of elephant trunks.

Animal and architectural motifs

Next to the animal world—and of more recent origin—architecture is the most important source of decorative motifs in furniture. In the late Middle Ages, the perpendicular tracery of Gothic architecture was transferred through the craft of the wood-carver to the fronts of chests. Italian chests and walnut cupboards of the same period were modelled on the marble sarcophagi of classical antiquity, which are entirely architectonic in form. During the Renaissance and Baroque periods the column was introduced as a strikingly decorative frontal feature in the form of table legs and on cupboards. The fronts of very big, heavy cupboards particularly lent themselves to architectonic composition corresponding to the portals and gables of houses (Figure 54). At about the same time, the ornamental wealth of the Renaissance broke through in rosettes, cupids, and fruits on panelling and frames.

During the Court Baroque period under Louis XIV in France, the royal official style left its mark not only on ornate pieces of furniture but also on panels, doors, mir-

By courtesy of the Victoria and Albert Museum, London; photograph, John Webb

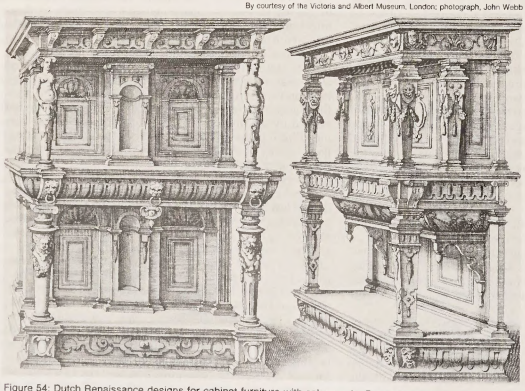


Figure 54: Dutch Renaissance designs for cabinet furniture with columns, by Paul Vredeman de Vries. (1567–1630?).

ror frames, and, indeed, even on the facades of palaces and châteaus and the layout of formal gardens. The coherence between interior and furniture was even more pronounced during the Rococo period and under Louis XVI, culminating temporarily in the furniture and rooms of the French Empire style.

The 19th century often seems to have offered nothing more than a breathless repetition of this coherence between the ornamental design of furniture and the architecture of the interior—both revivals of the styles of the past. A new style did not arise until the close of the century. French Art Nouveau furniture, with its gliding vegetable forms, must be seen in conjunction with the houses and rooms for which it was executed. The furniture of Antonio Gaudí, a Spanish architect and designer, for example, had a profound coherence with his own buildings; and the strangely expressive and stylized furniture of a Scottish architect, Charles Rennie Mackintosh, forms an integral part of his buildings and interiors in Glasgow.

Art Nouveau furniture

The influence of architecture on furniture can also manifest itself in a lack of ornament. There is a relationship, for example, between functionalistic architecture as it was first manifested in the 1920s at the Bauhaus in Germany and steel furniture designed by the German architect Mies van der Rohe.

KINDS OF FURNITURE

Chair. Of all furniture forms, the chair may be the most interesting. While most other forms (except the bed) are intended to support objects, the chair supports man. The term chair is used here in the widest sense, from stool to throne to derivative forms such as the bench and sofa, which may be regarded as extended or connected chairs, and whose character (*i.e.,* whether they are intended for sitting or reclining) is not clearly defined.

The social history of the chair is as interesting as its history as an art and craft. The chair is not merely a physical support and an aesthetic object; it is also an indicator of human worthiness. One is offered a chair; one does not just sit down anywhere at all without being asked to do so. Chair forms may also involve an indication of rank. At the old royal courts there were social distinctions between sitting on a chair with arms, on a chair with a back but no arms, and having to make do with a stool. In the 20th century, the director's or manager's chair has been an indicator of superior dignity, and even in democratic parliaments the speaker sits on a raised level.

As a furniture form, the chair encompasses a wealth of variations. There are chairs designed to match man's age and physical condition (the high chair, the wheelchair) and for his position in society (the executive chair, the throne). In the olden days there were chairs to be born in (birth chairs); in the 20th century, there have been chairs to die in (the electric chair). There are chairs with one, two, three, and four legs, chairs with or without arms, and chairs with or without backs. There are chairs that can be folded up, chairs on wheels, and chairs on runners.

Modern living has developed special chairs for automobiles and aircraft. All of these chair forms have been evolved to conform to changing human needs. Because of its close association with man, the chair appears to its full advantage only when in use. Whereas it makes no difference to one's appreciation of a cupboard or a chest of drawers whether there is anything inside or not, a chair is best seen and evaluated with a person sitting on it, for chair and sitter complement one another. Thus the various parts of a chair have been given names corresponding to the parts of the human body: arms, legs, feet, back, and seat.

Because the basic function of the chair is to support man, its value is judged primarily on how well it fulfills this practical role. In the construction of a chair, the designer is bound by certain static laws and principal measurements. Within these limits, however, he has great freedom.

The history of the chair covers a period of several thousand years. There are civilizations that have created distinctive chair forms, expressive of the highest endeavour in the spheres of technique and aesthetics. Among such cultures, special mention must be made of ancient Egypt and Greece; China; Spain and Holland in the 17th century; England in the 18th century; and France in the 18th century during the reigns of Louis XV and Louis XVI.

Egypt. Two ancient Egyptian chair forms, both the result of careful design, are known from discoveries made in tombs. One of these is a four-legged chair with a back, the other a folding stool. The classical Egyptian chair (Figure 55) has four legs shaped like those of an animal, a curved seat, and a sloping back supported by vertical stretchers.

The classical Egyptian chair

Photo by F.L. Kenett © George Rainbird Ltd. 1963



Figure 55: Golden throne from the tomb of Tutankhamen at Thebes, wood overlaid with gold and inlaid with faience, glass, and calcite, Egypt, 18th dynasty, c. 1350 BC. In the Egyptian Museum.

In this way a strong triangular construction was obtained. There was apparently no marked difference between the construction of Egyptian thrones and chairs for ordinary citizens. The main difference lies in the decorative ornamentation, in the choice of costly inlays. The Egyptian folding stool probably was developed as an easily portable seat for officers. As a camp stool the form persisted until much later times. But the stool also took on the character of a ceremonial seat, its mechanical function as a folding stool being forgotten. This can already be observed, from as early as 1366–57 BC in two stools, executed in ebony with ivory inlay work and gold mounts, from the tomb of Tutankhamen. They are in the form of folding stools but cannot be folded as the seats are of wood. The simple construction of the folding stool, consisting of two frames that turn on metal bolts and support a seat of leather or fabric fastened between them, reappears somewhat later in the Bronze Age folding chairs of Scandinavia and northern Germany. The best known of these is the folding stool, made of ashwood, found at Guldhøj (National Museum in Copenhagen).

Greece and Rome. The typical Greek chair, the klismos, is known not from any ancient specimen still extant, but from a wealth of pictorial material. The best known is the klismos depicted on the Hegeso Stele at the Dipylon burial place outside Athens (*c.* 410 BC) (Figure 52). It is a chair with a backward-sloping, curved backboard and four curving legs, only two of which are shown. These unusual legs were presumably executed in bent wood and were therefore subjected to great pressure from the weight of the sitter. The joints fastening the legs to the frame of the seat are therefore very strong and clearly indicated.

The Romans adopted the Greek chair; a number of statues of seated Romans show several examples of a heavier and apparently somewhat more crudely constructed klismos. Both types, the light and the heavy, were revived during the Classicist period. The klismos chair is found in French Empire furniture, in English Regency, and in special forms of considerable originality in Denmark and Sweden around 1800.

China. The ancestry of the chair in China cannot be traced as far back as in Egypt and Greece. Since the T'ang dynasty (AD 618–907) an unbroken series of drawings and paintings has been preserved showing the interiors and exteriors of Chinese houses and their furniture. Also preserved since the 16th century are a number of chairs of wood or lacquered wood that bear an astonishing resemblance to representations of older chairs.

Two major Chinese chair forms

As was the case in Egypt, there were two major chair forms in China: a chair with four legs and a folding stool. The four-legged chair is found both with and without arms but always with a square seat and straight stiles (upright side supports) to support the back. In one form, however, the stiles are slightly curved above the arms so as to conform to the rhythm of the S-shaped back splat (the central upright of a chairback). All three parts are mortised into the yoke-like top rail. While the design of the back splat exercised an influence on English chairs of the Queen Anne period, wooden members that only to a limited extent reinforce corner joints (and are loose into the bargain) represent a feature exclusive to Chinese chairs. The four legs pass through the seat frame, which closes about the rounded staves. All members are round in section or have rounded edges—it is as though one can just discern a bamboo tradition hovering in the background. The seat is uncomfortable and may have a plaited bottom. These chairs must have required the sitter to remain stiff and upright; for if too much pressure is exerted on the back, the chair has a tendency to topple over. In patriarchal Chinese homes of this period armchairs presumably were reserved for the senior members of the family, for they were held in great esteem.

The Chinese folding stool is presumed to have travelled to China from the West. It does not differ so very much from the Egyptian or Scandinavian folding stools, but it has a variation in that the top rail is elegantly joined to the two legs of the stool by means of a curved member, which is often provided with metal mounts. From a Western viewpoint the overall effect of both these furniture forms is stylized. The constructive and decorative elements are combined in a manner that is simultaneously naïve and refined. The pieced-together appearance is a result of the fact that the individual members do not appear to have been joined together with either glue or screws, but have been mortised into one another and locked into position in the manner of a Chinese puzzle.

Spain: 17th century. The Golden Age of Spain during the 17th century also left its mark on the chair. Paintings show a type of chair with a relatively crude wooden frame; a back and seat, nailed on, consisting of two layers of leather, with horsehair stuffing in between, stitched to produce a pattern of small pads. The front board and a corresponding board at the back could be folded after loosening some small iron hooks. Thus the chair was an easily portable piece of furniture for travelling which, at the same time, had the dignity of a four-legged, high-backed armchair.

Holland: 17th century. A low, square, upholstered type of chair can be seen in engravings of interiors of affluent Dutch homes by Abraham Bosse, a French artist (Figure 56), and in paintings by the Dutch artists Jan Vermeer and Gerard Terborch. Although this kind of chair is also found in countries where Dutch styles of interior decoration and Dutch furniture won favour, it is not certain that the form actually originated in Holland. Normally, the legs of the chair are smooth, round in section, and of slender dimensions; they are sometimes balustershaped (vase-shaped) or twisted. It is clearly a bourgeois piece of furniture and was made in considerable numbers, as can be seen from one of Abraham Bosse's engravings, in which a whole row of such chairs has been lined up against a wall. The form asserts itself by virtue of its harmonious proportions and fine upholstery in gilt leather or fabric bordered with fringes.

France and England: 17th and 18th centuries. The French Rococo chair in its most mature form—that is, as developed in Paris around 1750—spread over most of Europe and has been imitated or copied into the mid-20th century. The model owes its popularity to a combination of comfort and elegance. The seat conforms to the human body and permits a relaxed sitting position. The back is bow-shaped, the legs curved. Normally the seat and back are upholstered, and there are small upholstered pads on the armrests. Smooth transitions achieved between seat frame, legs, and back disguise all the joints, which are solidly constructed on craftsmanlike principles despite the absence of stretchers between the legs.

French Rococo chair

French Rococo chairs and imitations thereof employ wood of fairly thick dimensions; but all members are deeply molded, all superfluous wood has been cut away, and finer examples may be further embellished with very delicate and decorative carving. The wood may be left in

By courtesy of the trustees of the British Museum; photograph, J.R. Freeman & Co. Ltd.

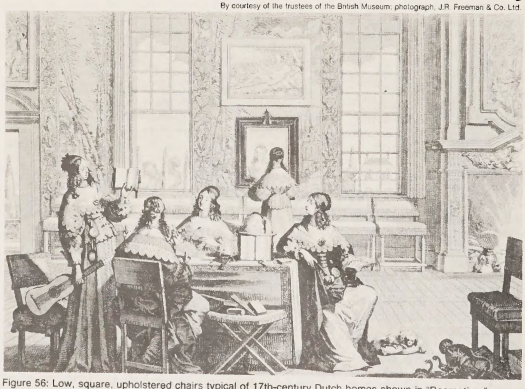


Figure 56: Low, square, upholstered chairs typical of 17th-century Dutch homes shown in "Recreation," engraving by Abraham Bosse, 1635.

its natural state, painted, or gilded. Silk damask or tapestry is used for the upholstery on the seat, back, and armrests; canework is sometimes used in place of upholstery.

English chairs of the 18th century are more differentiated in design than the French. The French taste for stylistic uniformity, which spread from the most distinguished circles in Paris and Versailles over most of France and won favour in several parts of the Continent, had no parallel in England. Prior to 1740, the most commonly used wood was walnut; thereafter, and for the rest of the century, it was mahogany. Walnut, though beautiful in hue, was soft and therefore less suited to wood carving than to rounded, curving forms. Outer surfaces, such as the back and seat frame, were usually veneered. During the walnut period, highly overstuffed armchairs, covered with leather or embroidered material, were also developed. The best upholstery of this period is precisely and firmly modelled and accentuated by braiding or tacks. When imports of mahogany became common, no specifically new chair designs appeared, but the character of the woodwork changed. Mahogany, having a firmer, closer grain, could be cut thinner, which meant that individual parts of the chair could be more slender in shape. Mahogany also lent itself better to carving than walnut. Carving was concentrated more on the arms and back than on the legs, which as a rule were straight and smooth with chamfered (bevelled) edges and molding. There was a wealth of variety in chairback designs, featuring elegant, pierced, vase-shaped splats or two upright posts connected by horizontal slats (ladderback).

Stick-back chair

Alongside the French Rococo chair and the best English chairs in walnut and mahogany, the more countrified stick-back chair was relatively unaffected by the stylistic changes of the day. Originally a medieval form, known, for example, from paintings by Pieter Bruegel the Elder and still found in mid-20th century in the churches and inns of southern Europe, the stick-back chair (in all of its variations) consists basically of a solid, saddle-shaped seat into which the legs, back staves, and possibly the armrests are directly mortised. This typically peasant form underwent a renewal and a process of refinement in England and America during the 18th century. Under the name Windsor chair (a term that seems to have been used for the first time in 1731) or Philadelphia chair, it became well-known and was widely distributed throughout the world. Related in form to the Windsor chair is an American chair made by members of the Shaker sect in the workshops they set up in their small, closed religious communities. Some of the best examples of these were executed around the middle of the 19th century at the Shaker headquarters in Pennsylvania.

Late 18th to 20th century. During the Neoclassical period, no basic changes took place in chair forms, but legs became straight and dimensions lighter. Backs in the shape of classical vases replaced the fanciful outlines of the Rococo period. Around 1800, freely executed imitations of Greek and Roman chairs of the klismos type, with curved legs and backrest, appeared. French chairs of the Empire period, executed in dark mahogany and embellished with ornate bronze mounts, created a ponderous effect.

In cheaper versions of inferior workmanship, bourgeois chairs of the 19th century carried on the traditions of the 17th and 18th centuries. The only real innovations were the bentwood (wood that has been bent and shaped) chairs in beech that became popular all over the world and are still made in the 20th century (Figure 57). Around 1900 the continental Art Nouveau and *Jugendstil* styles

By courtesy of the Technisches Museum fur Industrie und Gewerbe, Vienna



Figure 57: Bentwood rocking chair designed by Michael Thonet, Vienna, *c.* 1860. In the Technisches Museum für Industrie und Gewerbe, Vienna.

(French and German styles characterized by organic foliate forms, sinuous lines, and non-geometric forms), and the Arts and Crafts movement in England (established by the English poet and decorator William Morris to reintroduce standards of medieval craftsmanship), gave rise to original chair designs by Eugène Gaillard in France, Henry van de Velde in Belgium, Josef Hoffman in Austria, Antonio Gaudí in Spain, and Charles Rennie Mackintosh in Scotland. These new furniture styles did not exercise wide, let alone decisive, influence. The Art Nouveau chairs designed by the French architect Hector Guimard, for example, are collector's pieces, but his name is known to a broader public only because of his fanciful entrances to the Paris Métro.

Modern. After World War I, the Bauhaus school in

By courtesy of (left, centre) the Museum of Modern Art, New York (left) gift of Herbert Bayer, (centre) gift of Knoll Associates, Inc., (right) Herman Miller Furniture Co



Figure 58: *Twentieth-century chair design.*
(Left) Chrome-plate tubular steel armchair with canvas seat back and armrests, designed by Marcel Breuer, Germany, 1925. In the Museum of Modern Art, New York. (Centre) Chrome-plated steel lounge chair (Barcelona chair) with leather cushions and supporting straps, designed by Ludwig Mies van der Rohe, Germany, 1929. In the Museum of Modern Art, New York (Right) Molded plastic armchair reinforced with glass fibres, designed by Charles Eames, U.S., 1949.

Germany became a creative centre for entirely revolutionary thinking, resulting, for example, in tubular steel chairs designed by the architects Marcel Breuer, Ludwig Mies van der Rohe, and others (Figure 58). During World War II, the aircraft industry accelerated the development of laminated wood and molded plastic furniture. The dominant chair forms of this period go back to designs by a Finn, Alvar Aalto, Brun Mathsson, and an American, Charles Eames (Figure 58). Rapid technical developments, in conjunction with a severance from all tradition, suggest that completely new chair forms will probably be evolved in the future.

Table. *Fixed and mechanical tables.* In general, tables can be divided into fixed and mechanical types. The fixed table, consisting of a square or round top supported by one or more legs, is the least complicated from the viewpoint of craftsmanship. It is a form that requires wood of thick dimensions in order to make the joints by which the top is fastened to the legs strong enough to resist lateral pressure. Old Spanish or Italian tables are often constructed with sloping stretchers to counteract this pressure. The simplest way to make a table steady without exaggerating the dimensions of the individual parts is to fasten the legs to an underframe. Fixed tabletops can also make

By courtesy of the Kunstindustrimuseet, Copenhagen



Figure 59: Walnut table with wrought-iron stretchers, Spain, early 17th century. In the Kunstindustrimuseet, Copenhagen.

Pedestal table

do with a single leg; for example, the so-called pedestal table, terminating in a tripod or quadripod. Pedestal tables topple over easily, however, unless both top and pedestal are particularly heavy. Three-legged tables with a fixed top provide a more reliable support than a single-legged type but are unstable when subjected to uneven pressure from above.

The term mechanical refers to all tables whose tops can be enlarged or reduced according to need. Such tables may require pivotable or collapsible legs to augment the strength of the top. A familiar solution to the extension of a tabletop is the so-called Dutch system, known since the 17th century from Dutch engravings and paintings, in which the extension leaves, when pulled, slide out on sloping runners. When the leaves have been fully extended, the top is lifted and then dropped into place. The table height remains the same. The construction demands great accuracy and skill on the part of the craftsman. There are also more complicated forms of extension tables with runners enabling the legs as well as the leaves to be drawn out; extra leaves can then be inserted.

Tables with flaps also are constructed to take up less space when folded away and can be variously made, either with flaps that are supported by brackets that swing out on hinges or on so-called gate legs. During the 18th century, England was a leader in the design of ingenious folding tables, especially card tables. In the gateleg card table, the top can be folded so as to occupy half the space, and when opened is supported by a leg that swings out like a gate. In another system, the square underframe can be extended to form a rectangular top, the two sides being divided by hinges. On modern card tables, all four legs can be folded up within the frame surrounding the top; when not in use, the tables can therefore be stored easily.

Gateleg table

Historical forms and styles. Round stone tables on low pedestal legs are known in Egypt from the time of the pyramids (*c.* 2700 BC). Egyptian limestone reliefs also show tables of normal height. Dating from the later dynasties, crude wooden tables with archivectonic molding have been preserved. No tables have survived from ancient Greece. From the Roman ruins of Pompeii and Herculaneum, however, there are examples of monumental table supports or side members made of marble decorated with relief work and metal tables (Figure 64), many of them of the folding type. All wooden furniture has been lost, however.

Several wooden-topped communion tables dating from the early Middle Ages still stand in churches, hidden by altar cloths or built into boxes. Usually, such tables rest either on solid masonry or on a stone socle (a projecting member beneath the base of a superstructure), but they are sometimes elegantly supported by several columns. Generally, communion tables are made of stone, and since one stands before them, they are higher than the usual table. Examples of wooden tables preserved from the late Middle Ages are, as a rule, long narrow tops fastened to side members.

The interesting feature about the tables of the Renaissance and Baroque periods is their constructive and aesthetic design. Their thick and heavy tops rest on an underframe; the legs are baluster-shaped or turned, with deeply carved bulbous decoration. In the 17th century and later, table forms were widely differentiated and made for a great variety of purposes; *i.e.,* dining tables, library tables, drawing-room tables, card tables, tea tables, small candlestick tables, sideboards, and console tables.

Ming forms

From the Ming dynasty and the 18th century, several interesting Chinese fixed-top table forms have been preserved, in which the constructive elements are in some cases emphasized and in others deliberately disguised. Like other Chinese furniture forms, the tables create a stylized effect, with a naïve, calculated character. Chinese tables may be completely covered with lacquer and gilt ornamentation, but sometimes the wood is left in its natural colour.

Bed. In Homer's *Odyssey* there is a description of how Odysseus made his own bed: the trunk of an olive tree was cut to the exact shape and planed smooth; after holes had been drilled in the framework, oxhide thongs, dyed crimson, were threaded back and forth to make a pliant web; finally, the wood was embellished with inlay work in gold, silver, and ivory.

As a furniture form, the bed is as old as the chair. In principle the construction of the bed is extraordinarily simple: it consists merely of a rectangular platform raised in some way or other slightly above floor level. A considerable number of bed forms cannot be classed as furniture at all. Alcoves and bunks in ships, railway carriages, and airplanes belong more to the sphere of building trade joinery than to cabinetmaking.

That a number of beautiful and original bed forms of fine artistic execution have been created since antiquity is attributable to the fact that the bed gives the furniture designer rich possibilities in terms of framing and presentation, particularly in conjunction with textiles. Apart from the actual bedclothes, which will always be of greater importance than the actual platform and the surrounding framework, imaginative experiments combining the practical and the impressive—in four-poster beds and tentlike canopies, for example—have been made for centuries.

An Egyptian bier dating from the 1st dynasty (*c.* 3100–2890 BC) shows the original form of the bed: a rectangular framework of staves, round in section and mortised into one another so as to leave the ends free lengthwise, supported on four small legs carved to represent stylized lions' feet. Amusingly, the feet face in the same direction—as if they were walking with the dead person. This is characteristic of all Egyptian beds. Made of cedarwood, the light framework is higher at the head than at the foot; and whereas the foot is always terminated by a footboard, there is no board at the head. The beds were so constructed because the Egyptians when sleeping or resting used a stool-like support for the head (Figure 62). Essential to the Egyptian bed, countless examples of this piece of equipment—made usually of wood but sometimes of

ivory and faience—have been found in Egyptian tombs. The actual framework of the bed was often covered with plaited leather thongs.

In China, a bed in the form of a complete little house, with an anteroom in the form of a veranda, was placed in the middle of the room.

Before central heating and a knowledge of hygiene became common, the closed bed was the generally accepted form in cold climates. The simplest way to avoid drafts was to place the bed in an alcove—as was the practice in farmhouses right up to the 19th century. The most frequently encountered form of bed in European civilization, however, was the four-poster. Throughout the Middle Ages and later, the four-poster was developed in a variety of forms. Already during the Middle Ages, beds were designed for clearly ceremonial effect. The four posts supported an expanse of cloth that extended from the head like a canopy, just as the most distinguished row of choir stalls in a church was crowned by a baldachin (an ornamental structure resembling a canopy). Miniatures in illuminated manuscripts of the same period show tentlike beds entirely closed by drapery and curtains.

The four-poster

In the time of the absolute monarchies in the 17th and 18th centuries, pompous four-posters were developed in which the surrounding textile drapery completely concealed the wooden construction of the bed, thereby achieving a synthesis of practical and ceremonial considerations. Every palace or mansion had a chamber of state among its official reception rooms. Contemporary memoirs describe the complicated ceremony that took place at Louis XIV's daily awakening. Where his royal highness spent the night was his own concern, but his awakening was an act of state, in the conduct of which princes of the blood, dukes, and distinguished courtiers all had their respective duties: one would draw aside the bed-curtain, another would have the royal dressing gown ready, another the royal slippers. It was the first audience of the day, the king's levee. A large number of 17th- and 18th-century four-poster beds are still preserved in palaces, country houses, and museums; and most of them have a clearly dramatic, almost theatrical effect (Figure 60). The four-poster beds of the Baroque and Rococo periods, moreover, reflect great artistic refinement, especially in the rare instances in which they can still be seen in their original interiors complete with their entire textile adornment. Such beds of state are typical of continental Europe. In England and America, particularly toward the end of the 18th century, greater interest was taken in showing off the bedposts and the upper framework connecting them. Many English four-posters have slender, finely carved mahogany posts, whereas on the Continent the corresponding parts may be entirely covered with the same silken material as that used for the curtains, canopy, and bedspread.

During the Empire period in France an entirely new form of bed was developed and won favour throughout most of Europe. The design was inspired by the Roman couch as known from reliefs and from excavations in Pompeii and Herculaneum. The frame was very high, and the bed ends consisted of volutes (spiral or scrollshaped forms) of equal height. The bed was crowned by a tentlike superstructure, and the martial aspect was further emphasized by the use of spears to support the draperies and curtains; the whole bedroom, in fact, might well be draped like a tent (Figure 40). In these surroundings, the army commanders of Napoleon's time could feel like the caesars and consuls of ancient Rome. During a campaign, however, collapsible iron camp beds were more practical. Napoleon owned several and died in one on St. Helena in 1821. As a furniture form, the iron bed was a neutral framework built to support bedclothes and equipped with stanchions (upright supports) for curtains; it was light, transportable, and spartan.

Empire bed

Among plantation owners in the West Indies and the southern United States, a type of four-poster popular at the beginning of the 19th century was dominated by wood, rather than textile hangings. The posts supported very light, roughly made wooden frames, to which thin, white mosquito netting was fastened to protect the sleeper. The monumental and dignified effect was obtained by the quality of the woodwork. Of thick dimensions, the wood is

By courtesy of the Victoria and Albert Museum, London



Figure 60: Japanned, four-poster bed with canopy in the Chinese style, probably made by the firm of William Linnell for the Chinese bedroom at Badminton House, Gloucestershire, England, c. 1750–54. In the Victoria and Albert Museum, London.



Figure 61: Renaissance cassone, painted and gilded wood, Florence, 15th century. In the Victoria and Albert Museum, London.
By courtesy of the Victoria and Albert Museum, London; photograph, John Webb

solid mahogany polished to a high gloss. The four bedposts are not necessarily identical at the head and foot of the bed, but all have bulbous and turned sections, exaggerated almost to the point of crudeness. The headboards and footboards are imaginatively designed with voluted gables (triangular decoration) and galleries (ornamental railings) supported on pillars. Besides the practical function of these West Indian beds, they also served to indicate the importance of their owner; like the royal four-poster of the days of absolute monarchy, they clearly showed the difference between master and slave.

In the 20th century, the bed has belonged exclusively to one's private life; and compared with those of the past, modern beds are simple. Four-posters are still "modern," possibly because they appeal to something primitive, namely the sensation of sleeping in a tent. In general, development has been concentrated on improving the quality of bedclothes and increasing the amount of comfort by attention to springs, spring mattresses, eiderdowns, and pillows. The actual woodwork of the bed is usually restricted to joined veneered sections of laminated board, canework sometimes being used for the headboards and footboards.

Storage furniture. *Chest.* The chest, including the coffin (and sarcophagus), is an ancient primitive furniture form that has survived into the 20th century. The design of a clothes chest is optional; its size depends on changing demands. The construction of a coffin, on the other hand, is a set task. The format is determined by certain principal dimensions, and the human figure has at all times exercised an influence on the shape of the coffin; the Egyptian mummy case, which takes on the form of the swathed corpse, is an example. Traditional features of ancient Roman sarcophagi, and simplified versions of the monumental style of the Baroque and Renaissance periods continue to thrive.

Early medieval chests

The principal constructional features of early medieval chests lasted until the Renaissance. The so-called Oseberg ship, dating from the Viking era (9th century AD) and discovered in 1904 in Vestfold, Norway, included among the furniture on board a chest made of oak planks secured by iron bands. The planks are not mortised together, and the end sections stand vertical, thereby forming feet, wider at the bottom than above. The lid is formed by a single curved oak plank that has been roughhewn into shape. The bottom of the chest rests in a groove cut into the end sections. The wooden construction, a primitive form of carpentry, is held together by broad iron bands, the nails are tin-plated. In this Oseberg chest, the iron mounts essential to the construction constitute the decorative element as well. All medieval chests are developments of the same principle: a piece of carpentry with decorative iron mounts, but the principle found freer application in medieval church doors than in the chests of the period.

The chest often appears in portable form as a traveller's trunk that can also serve as a stationary piece of furniture. A number of painted, parchment-covered Florentine chests dating from the middle of the 15th century have been preserved. These were used as trunks by young girls on their way to enter a convent and later stood in their cells as pieces of storage furniture for clothes and other personal belongings. A "nun's chest" of this type is in principle quite different from the sumptuous cassoni of the Italian Renaissance that were adorned with gilded stucco work and painted panels (Figure 61). Cassoni were stationary pieces of palace furniture. Specifically designed for travelling, however, were Javanese camphorwood chests that made the long voyage round the Cape of Good Hope full of stuffs and spices and eventually came to rest in an English manor house or in a gabled Dutch mansion in Amsterdam. The plank construction with metal mounts is of primitive craftsmanship. The large, smooth expanses of reddish-brown wood, with their elaborate openwork brass mounts and big, chased bolt heads to take the brunt of rough handling, have a kind of sophisticated crudeness about them. On later camphorwood chests the brass mounts are sunk flush with the surface of the wood, just as on portable writing desks and toilet cases of the French Empire period. Veneered wood was not suitable for chests intended for travel purposes, but it was possible to cover the entire chest with leather fastened with metal nails, possibly to form a pattern. Several beautiful, leather-covered chests made in Italy and Spain in the 17th century are known, and the form persisted in the large wardrobe trunks of succeeding centuries.

"Nun's chest"

When furniture-making techniques demanding the skill of the cabinetmaker evolved during the Renaissance, frames, panels, and carving appeared on chests. In southern Europe, walnut lent itself admirably to carving; in northern Europe, oak. While the Italians were inspired by the molding and decorative plant ornamentation of the stone sarcophagi of ancient Rome, in northern Europe late medieval wood carving traditions were continued. As a rule the carved woodwork was picked out (trimmed) with paint and gilded. In the 18th century, the chest was largely supplanted for storage purposes by the chest of drawers and the commode (low chest of drawers), but it never entirely disappeared. Particularly in the big country houses of England and America, chests of mahogany or walnut were used for a long time, often having drawers in the bottom and finely fashioned brass mounts that revealed Chinese influence.

Cupboard. Strictly speaking, the cupboard is a derivative form of the chest. Early Renaissance cupboards resembled two chests placed one on top of the other, but they were opened from the front by means of doors. The design and construction of the cupboard's pronounced front have always provided ample scope for artistic composition, and it is no mere coincidence that the cupboard more than any other furniture form should have closer links with architecture. It literally invited an architectonic composition: socle, columns, cornice. This development can be traced from the close of the Middle Ages in a large number of southern German and Tirolean cupboards bearing late Gothic perpendicular tracery and smooth surfaces veneered with ashwood. Very large cupboards took

on their most striking form, however, during the Renaissance in 17th-century Holland and northern Germany. In molding and composition, they have much in common with architectural facades, but their picturesque and textural effects are the result of refined craftsmanship. The use of veneer was almost essential if these large expanses of wood were to be infused with life. A carcass of wood was given a veneer of fine walnut; socle, frames, columns, and cornice were decorated with veneered black ebony. The doors were furnished with strong locks, and the keyhole was concealed behind a sliding middle column. The cornice was often decoratively crowned with a set of Dutch faience or Chinese porcelain vases. These heavy cupboards were made to appear lighter by placing them on big, turned ball feet. In marked contrast to the European Baroque cupboards, Chinese cupboards of the same period were simple, smooth-surfaced, and boxlike. Their construction was based on a simple system of uprights and frames, and as a rule they were made in pairs. If painted, a large decorative painting was spread across the entire surface, including the doors. Inside, Chinese cupboards are finished with great care and painted in a different colour from the outside. The mounts are of various white and yellow metal alloys, smooth, either round or square; and the locks are secured with prismatically designed padlocks. Japanese and Siamese cupboards, apart from certain independent features, follow the old Chinese traditions.

Chinese cupboards

The clothes cupboard of the 19th and 20th centuries, an indispensable piece of bedroom furniture wherever there are no built-in cupboards, is based on traditional features of the 18th-century English clothespress but equipped to meet the changing fashions of modern times.

Bookcases. Bookcases or bookshelves are a less interesting form of storage furniture from the viewpoint of furniture history. Perhaps the most significant innovation appeared in 18th-century England in the bookcase with adjustable shelves and a closed-off lower section for folio files. The shelves were protected by glass doors consisting of an ingenious trelliswork of carved wood. Bookcases and shelves become interesting only when they form part of specially designed library interiors and when several shelves full of books create an intimate, compact whole.

Mixed forms. Apart from the kinds of storage furniture already mentioned, there are numerous combination forms. An ordinary table can be used as a writing desk, and the only differences between the typical French Rococo writing desk of the 18th century and other tables are the drawers in the underframe and the leather-covered top. The novelty of Louis XV's writing desk consists of a rolltop device for closing the writing flap. In England a special type of writing desk was developed which, besides drawers in the underframe, has a side cupboard fitted with additional drawers and, occasionally, sliding trays. Some have a false drawer front that can be pulled out to form a writing surface. When a writing desk has a cupboard built on the top of it and is placed on a chest of drawers, the result is a cabinet or secretary. There are also bookcases with lower sections equipped with a flap, either hinged or sliding, for writing. All of these combinations, frequently of ingenious design, were made anonymously in England during the 18th century, apparently having arisen from a desire on the part of the well-to-do middle classes to develop a sophisticated and differentiated pattern of life.

English writing desks

A special group of storage furniture embraces the various forms of corner furniture, low or high cupboards that were made in pairs (just as in the case of several other old furniture forms) particularly for small rooms, in which they became fixed components of the interior scheme.

Kitchen furniture and furnishings. Kitchen furniture and furnishings go back to antiquity. In the Middle Ages, the kitchen, with its fireplace, was the most centrally placed room in the home. Later, closed fireplaces were constructed in the form of stoves; and cupboards, sinks, and plate racks were fixed to the wall. The kitchen in a modern home, if not combined with a dining area, is a small room filled with equipment. On the other hand, institutional kitchens have expanded enormously. Outdoor cooking equipment, such as various forms of open-air grills, also forms part of modern kitchen furniture.

Bathroom furniture and fixtures. Bathrooms in large private homes were not unknown in the 18th century, and splendidly equipped marble bathrooms are still preserved in several European palaces and mansions. But not until the 19th century did bathrooms in private homes become more commonplace. Fixtures generally include a toilet, bidet, washbasin, bath, mirror, and shelves. In the 20th century the equipping of bathrooms became a separate industry with a wide variety of special forms of bathroom furniture and fixtures. The materials used are porcelain, enamel, plastic, wood, and stainless steel.

Specialized furniture. Office furniture in the widest sense of the term has undergone rapid developments since mid-19th century. Such pieces as the high desks used by clerks in old offices and the big American rolltop desks have been replaced by carefully designed standard forms of writing desks with side cupboards, typewriting tables, filing cabinets, and office chairs with adjustable backs and swivel seats. From office furniture one passes naturally to the vast sphere of institutional furniture: theatre furnishings in the form of rows of connected seats, restaurant furniture, furniture for conference rooms, laboratories, workshops, and factories. Several of these specialized furnishings reflect past traditions. The way in which the British House of Commons is furnished, for example, derives without doubt from the pattern in which choir stalls were grouped in medieval churches; whereas the semicircular, often amphitheatrically designed assembly halls of the United States Congress and the parliaments of many European countries are developed forms of academies of surgery or other university auditoriums. Similarly, museums, libraries, and archives have their special furniture in the form of showcases, desks, special tables, and socles.

Office furniture

There are also furnishings for movable premises, primarily railway carriages equipped for sleeping or dining, passenger ocean liners with cabins, airplanes, buses, coaches, and private cars.

Finally, there is the large, highly heterogeneous group comprising outdoor or open-air furniture; for example, furniture for gardens, balconies, terraces, and solaria.

KINDS OF ACCESSORY FURNISHINGS

Besides the aforementioned kinds of furniture and all the many special forms (which it would be almost impossible to list), there is an extensive group of accessory furnishings that is not furniture in the strict sense but, nevertheless, constitutes an important element in the furnishing of interiors. Included here are clocks and other mechanical works, mirrors, textiles, screens, stoves, and fireplaces; and a number of smaller articles made by cabinetmakers, such as boxes, caskets, sewing tables, wastepaper baskets, lighting fixtures, frames, panelling, and floor surfaces.

Clocks. Clocks are considered furnishings if the movement is enclosed within a case, which need not necessarily be of wood. Clocks can be divided into table clocks and long-case clocks. There were two creative centres for table clocks, namely England and France. In 17th- and 18th-century France, the table clock became an object of monumental design, the best examples of which are minor works of sculpture. The actual movement is framed by a marble socle, and the clockface by a sculptural frame of solid bronze incorporating freely molded figures and ornamentation. Some of France's best sculptors and bronze casters were engaged in the creation of decorative frames for clock movements. A French speciality, imitated elsewhere on the Continent, was the wall clock, or so-called cartel clock, the earliest examples of which were designed by a goldsmith and ornamentalist, Juste Meissonnier. The clockface is the centre of an ornament, or rocaille-cartouche, cast in bronze, sometimes garnished with figures of symbolic significance; for example, Time, a man with a scythe, or a crowing cock. In England, where tastes were more bourgeois, the fine movements made by skillful London clockmakers were built into wooden cases, architectonic in composition and featuring pilasters (partly recessed columns) and cornices. Simple walnut cases could be adorned with metal ornaments and brass balls. The more expensive table clocks were concealed in cases embellished with inlaid wood or tortoiseshell.

Table clocks

Long-case clocks

Long-case clocks were also made in France and England. French long-case clocks are monumental and richly designed. In the reign of Louis XIV there were long-case clocks of the boulle type with metal and tortoiseshell inlay work. Later, in the 18th century and especially during the Rococo period, the case that concealed the weights acquired more dramatic form; richly inlaid wooden surfaces were framed and adorned by magnificently gilded Rococo ornaments in bronze. The English long-case clock was to a greater extent a piece of furniture, and the main features of its construction remained unaltered throughout the 18th century. The longcase clock stands on a base, or socle, from which the somewhat narrower case for the weights rises up, crowned by the framework of the actual movement and clockface. The last-named section is in reality a table clock mounted on a weight case. Each individual section of the long-case clock is thus clearly separate; each has its distinct function; and no attempt was made, as in France, to veil the independence of the individual parts. The weight case is provided with a door in which there may be a window through which the position of the weights can be observed.

During the 18th century, barometers became increasingly popular. The mechanism was provided with a decorative wooden framework intended to harmonize with the other furniture in a room.

Mirrors. The use of mirror glass in furnishings arose during the 17th century. The discoloration of the melted glass because of silvering and, not least, the prohibitive cost and difficulty of manufacturing mirror glass of considerable size restricted the possibilities of large-scale application. The mirror gallery at Versailles (Figure 27) was thus an outstanding technical achievement for its time. When Louis XIV strode through the gallery at the head of his court, the glass walls reflected the diamonds in his crown. This effect was imitated to a greater or lesser degree in all the courts of Europe. In the 18th century the wall mirror found its way into most interiors. The popularity and wide distribution of mirror glass was stimulated by the need for an increased amount of artificial light. During the 16th and 17th centuries, this need had been satisfied by placing candles in front of highly polished concave metal plates. By using silvered mirror glass, the light effect was multiplied. From then on, large mirrors hung over console tables were a necessary and functional part of rooms illumined by artificial light.

Mirror gallery at Versailles

Fabrics. The use of fabrics in furnishing rooms is closely bound up with the need for heating. In the primitively heated rooms of the Middle Ages, textiles were used to keep out cold and drafts. In 12th- and 13th-century churches, painted textile drapery can still be discerned beneath the picture friezes. In rather cold churches, just as in poorly heated homes, loosely hung textile wall coverings were of the greatest importance. They were hung loosely because of the practice of taking them down and moving them, together with the relatively few items of furniture, according to need. It was not until the end of the 17th century and during the 18th century that tapestries and other forms of textile wall hanging became fixtures; that is, fastened to the wall within frames. Wall pictures made of paper and, subsequently, patterned wallpaper became a cheaper substitute for textile wall hangings during the 19th century. Screens or room dividers were often covered with textiles, partly to afford protection against direct radiant heat and partly to create cozy corners in large rooms. Framed screens were often covered with pieces of tapestry, with other woven materials, or with gilt leather. (See the section *Tapestry* below.)

Portable wall hangings

Fireplaces. The heating of rooms and large halls remained a major problem until the advent of modern central heating systems. The open hearth was replaced during the late Middle Ages by the fireplace, which is merely an architectonic way of framing the burning logs. During the period when it was important as a source of heat, the fireplace became the object of design work by significant artists. A Scottish architect, Robert Adam, and his brothers and an Italian architect and engraver, Giambattista Piranesi, made considerable artistic contributions to the design and construction of fireplaces.

Other accessory furnishings. Small utility objects constitute an important part of the furnishing of interiors. Several of them are the work of cabinetmakers; for example, boxes for writing paper and playing cards, caskets for letters and documents, trays for serving or presentation. Accessory furnishings include the various articles, large and small, that are employed in the course of domestic work—from small looms to lace pillows, spinning wheels, embroidery frames, and sewing tables. Women's chattels, partly in the form of equipment for domestic needs and partly in the form of items of storage furniture for such small items as pins, scissors, wool, and materials, all had their place in the home.

Finally, the structure and decoration of the walls, ceilings, and floors—for example, panelling, stucco work, parquet flooring, carpets—also come under the heading of accessory furnishings. (Er.L.)

History

WESTERN

Egypt. Beds, stools, throne chairs, and boxes were the chief forms of furniture in ancient Egypt. Although only a few important examples of actual furniture survive, stone carvings, fresco paintings, and models made as funerary offerings present rich documentary evidence. The bed may have been the earliest form; it was constructed of wood and consisted of a simple framework supported on four legs. A flax cord, plaited, was lashed to the sides of the framework. The cords were woven together from opposite sides of the framework to form a springy surface for the sleeper. In the 18th dynasty (*c.* 1567–1320 BC) beds sloped up toward the head, and a painted or carved wooden footboard prevented the sleeper from slipping down (Figure 62).

Construction of the Egyptian bed

The great beds found in the tomb of Tutankhamen were put together with bronze hooks and staples so that they could be dismantled or folded to facilitate storage and transportation; furniture existed in small quantities and when the pharaohs toured their lands, they took their beds with them. In the same tomb was a folding wooden bed with bronze hinges.

By courtesy of the Museum of Fine Arts, Boston

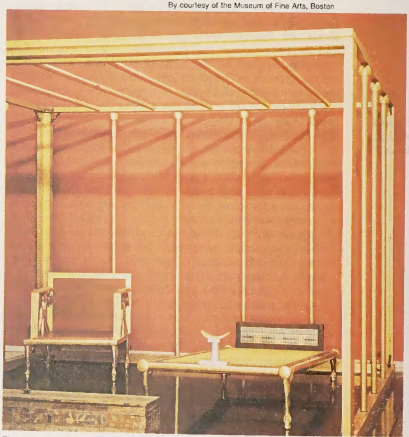


Figure 62: Reconstructed bed canopy with bed, chair, and curtain box from the tomb of Queen Hetepheres at Giza, wood overlaid with gold foil, Egypt, 4th dynasty, *c.* 2600 BC. A bed, with detachable footboard of inlaid faience and silvered headrest, and a low armchair stand beneath a canopy designed to take hangings. In the Museum of Fine Arts, Boston. Original in the Egyptian Museum.

Instead of pillows, wooden or ivory headrests were used. These were so essentially individual, being made to the measure of the owner, that they were often placed in tombs to be used by the dead man on his arrival in the land of eternity. Folding headrests were probably for the use of travellers.

Early stools for ceremonial purposes were merely squared blocks of stone. When made of wood, the stool had a flint seat (later shaped concavely) covered with a soft cushion. In time the stool developed into the chair by the addition of a back and arms. Such throne chairs were reserved for use by personages of great importance. Footstools were of wood. The royal footstool was painted with the figures of traditional enemies of Egypt so that the pharaoh might symbolically tread his enemies under his feet. Carvings of animal feet on straight chair legs were common, as were legs shaped like those of animals. Boxes, often elaborately painted, or baskets were used for keeping clothes or other objects. Tables were almost unknown; a pottery or wooden stand supporting a flat basketwork tray held dishes for a meal, and wooden stands held great pottery jars containing water, wine, or beer.

The Egyptians used thin veneers of wood glued together for coffin cases; this gave great durability. Egyptian furniture in general was light and easily transportable; its decoration was usually derived from religious symbols, and stylistic change was very slow.

Mesopotamia. The furniture of Mesopotamia and neighbouring ancient civilizations of the Middle East had beds, stools, chairs, and boxes as principal forms. Documentary evidence is provided chiefly by relief carvings. The forms were constructed in the same manner as Egyptian furniture except that members were heavier, curves were less frequent, and joints were more abrupt. Ornament was richly applied in the form of cast-bronze and carved-bone finials (crowning ornaments, usually foliated) and studs, many of which survive in museums. Mesopotamia originated three features that were to persist in classical furniture in Greece and Italy and thus were transmitted to other western civilizations. First was the decoration of furniture legs with sharply profiled metal rings, one above another, like many bracelets on an arm; this was the origin of the turned wooden legs so frequent in later styles. Second was the use of heavy fringes on furniture covers, blending the design of frame and cushion into one effect; this was much lightened by classical taste but was revived in Neoclassicism. Third was the typical furniture grouping that survived intact into the Dark Ages of Europe: the couch on which the main personage or personages reclined for eating or conversation; the small table to hold refreshments, which could be moved up to the couch; and the chair, on which sat an entertainer—wife, hetaera (courtesan), musician, or the like—who looked after the desires of the reclining superior personages. From this old hierarchy of furniture derive the cumbersome court regulations concerning who may sit and on what, persisting in the palaces and ceremonies of 20th-century monarchs.

Mesopotamian ornamentation

Greece. Principal furniture forms were couches, chairs (with and without arms), stools, tables, chests, and boxes. From extant examples, the depiction of furniture on vases and in relief carvings, and literary descriptions, much more is known about Greek furniture than about Egyptian. At Knossos, a built-in throne of stucco, much restored, is often considered to represent pre-Hellenic furniture in the Aegean area (Figure 17). Primitive Aegean pottery shows rounded chair forms, perhaps indicating basketry models, and Bronze Age sculpture shows complex-membered chair frames.

The Greek couch

In ancient Greek homes, the couch, used for reclining by day and as a bed by night, held an important place (Figure 63). The earliest couches probably resembled Egyptian beds in structure and possibly in style. The legs occasionally imitated those of animals with claw feet or hoofs, but usually they were either turned on the lathe and ornamented with moldings or cut from a flat slab of wood sharply silhouetted and decorated in various ways—with incised designs or with volutes, rosettes, and other patterns in high relief. From about the 6th century BC, the legs projected above the couch frame; these projections became headboards and footboards, the latter eventually made lower than the headboards. In Hellenistic times headrests and footrests were carved and decorated with bronze medallions carrying busts of children, satyrs, or heads of birds and animals in high relief. Turned legs largely replaced rectangular ones. Although a bronze bed of the 2nd century BC has been found at Priene and marble couches sometimes occur in tombs, the usual material was wood. The legs often terminated in metal feet and sometimes were encased in bronze moldings, and the rails also were sometimes covered with bronze sheathing.

Klismos and *diphros* chairs

From the Greek Archaic period onward many varieties of individual seats are known, the most imposing, perhaps, being elaborately adorned, high-backed ceremonial chairs of wood or marble. Like the couches, they were supported on turned legs, legs cut from a rectangular piece of wood, or legs with animal feet; they frequently had arm rails. Another type of boxlike seat with no feet and with or without a back is also found. The klismos chair was lighter and had a curved back and plain, sharply curved legs, indicating a great mastery of wood-working. The *diphros* was

By courtesy of the Staatliche Antikensammlungen und Glyptothek, Munich



Figure 63: Greek couch and rectangular table from an Attic amphora decorated by Andocides with scenes of Heracles dining, c. 510 BC. In the Staatliche Antikensammlungen und Glyptothek, Munich.

a stool standing on four crossed, turned legs, sometimes connected by stretcher bars and sometimes terminating in hoofs or claw feet. The convenience of folding stools was realized at an early date, and the *diphros* was popular.

Greek tables were usually small and easily portable. An interesting type had an oblong top supported by three legs, two at one end and one at the other. These legs usually tapered from the top and terminated in claw feet, and the bronze and stone examples which are occasionally found show carved flutings on the front of the legs and scroll ornament at the side below the table tops. Rectangular tables with four legs were also used, as were round tops.

Rome. Principal furniture forms were couches, chairs with and without arms, stools, tables, chests, and boxes. Excellent documentary evidence is found in mural paintings, relief carvings, and literary descriptions. Extant examples are more common than those of the ancient Near East: a wealth of bronze furniture was recovered at Pompeii; at Herculaneum even wood pieces were partly preserved.

The Roman couch

As in Greece, the couch was a principal furniture form. At Pompeii couches with bronze frames closely resembled Greek examples. Gold, silver, tortoiseshell, bone, and ivory were used for decoration, with veneer of rare woods. Later couches, found in Italy and in distant parts of the empire, were characterized by the high back and sides.

Roman chairs developed from Greek models. The Greek throne chair evolved into a small armchair with solid rounded back made in one piece with sides set on a rectangular or semicircular base. This armchair was often of wickerwork, wood, or stone. The Greek klismos chair was given heavier structural members by the Romans and was called the cathedra.

The Romans developed a decorative type of stool, often made in bronze. This was supported by four curved legs, ornamented with scrolls. The folding stool, with cross legs sometimes connected by stretcher bars, was used both by Roman officials and in households. Remains of folding stools are known from sites such as those at Ostia, Italy, and barrows in Britain—on the Essex-Cambridgeshire border, and in Kent. This developed into a stool that had more solid double curved legs; examples were found at Pompeii. An example in iron with bronze decorations, even heavier in form, was found at Nijmegen, in The Netherlands.

Roman tables

Tables with round and rectangular tops and three and four legs were common. Tables with round tops and three legs of animal form became increasingly popular from the 4th century BC onward (Figure 64). A nearly complete wooden table, found in Egypt and now in the Palais du Cinquantenaire, Brussels, is decorated with swans' heads with graceful necks rising out of a band of acanthus foliage, below which are very realistic antelope legs, with hoofs instead of claw feet. This type of table seems to have been popular throughout the Roman empire, as it often appears on tombstones depicting funerary banquets. It is known that citrus wood and Kimeridgian shale were favourite materials. Several complete tables found at Pompeii and Herculaneum, usually in gardens or open courts, are made of marble and decorated with beautifully carved heads of lions and panthers. Another type of smaller table is round or rectangular with only one central leg. Also found are pairs of solid slabs ornamented in high relief, carrying carved tops of marble or wood.

SCALA—Art Resource

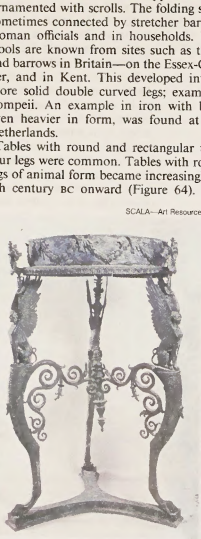


Figure 64: Roman table or stand with circular top from the Temple of Isis at Pompeii, bronze, before AD 79. The clawed legs, connected with elegantly scrolled braces, are surmounted by winged sphinxes. In the Museo Archeologico Nazionale, Naples.

Pompeian wall paintings show that plain, undecorated wooden tables and benches were used in kitchens and workshops, and some household possessions were kept in cupboards with panelled doors. Rectangular footstools, sometimes with claw feet, were used with the high chairs and couches. Small bronze tripods and stands were also items of Roman furniture. Clothes and money were stored in large wooden chests with panelled sides, standing on square or claw feet. Roman treasure chests were covered with bronze plates or bound with iron and provided with strong locks. Jewelry and personal belongings were kept in caskets, in small round or square boxes, or even in baskets.

Storage chests

Middle Ages. *Early Middle Ages.* With the collapse of the Roman Empire during the 4th–5th centuries, Europe sank into a period in which little furniture, except the most basic, was used: chairs, stools, benches, and primitive chests were the most common items. Several centuries were to pass before the invading Teutonic peoples evolved forms of furniture that approached the Roman standard of domestic equipment.

Comparatively little furniture of the medieval period in Europe has survived, and only a handful of these pieces date from before the end of the 13th century. One reason for this is the perishable nature of wood, but more important is the fact that furniture was made in relatively small quantities until the Renaissance. Much of the earlier history of furniture has to be drawn from contemporary literature, illuminated manuscripts, Romanesque and Gothic sculpture, and later inventory descriptions.

There is evidence that certain ancient traditions of furniture making, particularly that of turnery, influenced early medieval craftsmen. Turnery was used in making chairs, stools, and couches in Byzantium, and it seems that this technique was known across Europe as far north as Scandinavia. The Anglo-Saxon epic poem *Beowulf*, which gives some glimpses of the domestic economy of western Europe in about the 7th century, mentions no furniture other than benches and some kind of seat or throne for the overlord.

Later Middle Ages. In the 14th and 15th centuries there were many developments both in construction and design of furniture throughout Europe; a range of new types, among them cupboards, boxes with compartments, and various sorts of desks, evolved slowly. Most of the furniture produced was such that it could be easily transported. A nobleman who owned more than one dwelling place usually had only one set of furnishings that he carried with him from house to house. Anything that could be moved, and this frequently included the locks on the doors and the window fittings, was carried away and used to furnish the next house en route. Furniture was so scarce that it was quite usual for a visitor to bring his own bed and other necessities with him. These conditions had a double effect on medieval furniture, not only making it difficult for men to possess more than the basic types of furniture but also affecting the design of the furniture itself. Folding chairs and stools, trestle tables with removable tops, and beds with collapsible frameworks were usual.

The need for easily transported furniture

The religious houses were an exception to this in that they enjoyed a certain security denied to the outside world. Much of the best furniture of this period was therefore made for use in churches and monasteries, and many of the ideas and developments that were later to add to the domestic comfort of Europe originated in the cloister. An

example can be seen in the early development for ecclesiastical use of the various types of reading and writing furniture, such as lecterns and desks, that show ingenuity in construction. Throughout the Middle Ages and well on into the 16th and 17th centuries, all types of furniture remained scarce, and any reasonably good furniture belonged to the nobility and the wealthy merchants. The household equipment of the peasantry throughout Europe, even as late as the 18th century, was frequently crude in design and roughly constructed.

Framed panelling

Framed panelling had been used in ancient times, as examples found at Herculaneum testify; its reintroduction in the Burgundian Netherlands at the beginning of the 15th century was an improvement that soon spread throughout western Europe. Panelled construction solved the problem of building large surface areas, as on the front of a chest or cupboard, which before this time had been limited by the size of individual planks. These planks, usually hewn with an adz, were heavy and liable to warp and split. Panels could be cut thinner, the main strain being taken by the framework, and the furniture was therefore lighter; moreover, if the panels were not fitted too tightly in their stiles, the wood was less likely to split if it did warp. Now that it was possible to construct larger surface areas, a new range of storage furniture, cupboards and chests in particular, was developed.

Other constructional improvements of the 15th century included the introduction of drawers into cupboards and similar storage furniture, and neater and more efficient joints, such as the mitre and the mortise and tenon. Panelling was frequently decorated with a flat form of ornament called linenfold, or parchment. Linenfold was widely used in the north of France, Flanders, Low Germany north to the Baltic, Scandinavia, and England. The linenfold of France, the Low Countries, and Germany is carved with a sharper definition and greater delicacy than was usual in England and elsewhere. Both panelled furniture and room panelling were decorated with linenfold. Other forms of carved decoration on furniture became more common during the 15th century, when surfaces were carved with tracery and other Gothic motifs. During the Middle Ages a great many pieces of furniture, including those with carved decoration, were painted and sometimes gilded, a practice that continued well on into the Renaissance (the present state of existing pieces, with their plain wooden surfaces, is misleading). Chairs, tables, and various types of cupboards were also frequently draped with bright fabrics, while chairs, settles, and other seat furniture were provided with cushions.

The medieval chest

The chest was the basic type of medieval furniture, serving as cupboard, trunk, seat, and, if necessary, as a simple form of table and desk. It was from this versatile piece of furniture that several other types, such as the cupboard and the box chair, were evolved. Chests were made of six planks, crudely pegged or nailed together and frequently strengthened with iron banding. Examples of this sort, dating from the 13th century and in many instances found in churches, are among the earliest pieces of extant European furniture. The chest remained one of the most important pieces of furniture until the end of the 15th century, when on the Continent the cupboard began to compete with it in usefulness.

Chairs remained scarce throughout the Middle Ages, and occupation of a chair long symbolized authority or a mark of honour, and even a large house might possess only chairs for the lord and his wife and perhaps another for a distinguished visitor; the use of the word chairman is a modern reflection of this medieval custom. Early chairs constructed of turned spindles, seen in Romanesque sculpture, have already been mentioned. Later there were two main types. One was a variety of folding chair, with X-shaped frame, made of both wood and metal, the seat and back consisting of rectangular strips of some strong fabric or leather (Figure 65). Eventually there evolved a heavier type of chair. This was basically a development of the chest, and in many cases the seat was hinged, allowing the base to be used for storage. Panelling, often carved with linenfold and sometimes with other Gothic motifs, was used on the back, arms, and base. Many of these chairs had exaggeratedly high backs terminating in elaborately carved canopies; some were freestanding, while others had their backs fixed to the wall in the manner of a church stall. Settles were also used for seating during the 15th century. An innovation on the Continent was the settle with a pivoted bar forming the backrest, which could be swung over to allow a person to sit on either side—evidence of the weight of the furniture of this period.

Tables

Tables were mainly of trestle construction (with a braced frame serving as a support for the tabletop) with long rectangular tops that could be dismantled. During the 15th century on the Continent, smaller tables were made which could be more conveniently moved and, especially, drawn up to the fire. Various forms of cupboards, ambries, and dressoirs were developed at this time, panelled and decorated with linenfold or Gothic carved ornament. All these types were basically a chest with doors, of simple rectangular form raised on legs; elaborations of construction and decoration soon followed, as did the specialization of their functions. Cupboards, dressoirs, and credence (sideboard or buffet) tables were used for the storing of plate and for serving at banquets, the plate being displayed on the top and on shelves above and below the main serving surface. Top shelves were sometimes cantilevered or projected on brackets to free the front corners of this surface for use. Other cupboards were made to hold food and day-to-day provisions; in the case of food, or dole cupboards as they were called, the front and sides were pierced for ventilation.

Medieval beds are known from documents and a few late examples. Recalling Egyptian beds, throughout most of this period a diagonal surface, lifting the head high, was common. Some beds had daringly cantilevered ceilings supported from the headboards.

By courtesy of the Metropolitan Museum of Art, New York, The Cloisters Collection



Figure 65: Massive limestone mantelpiece from Alençon and chairs with X-shaped frames, France, 15th century. In the Metropolitan Museum of Art, New York, The Cloisters Collection.

English furniture

Little English furniture survives from medieval times, and, as on the Continent, information must be sought in contemporary references and from the picture of domestic interiors in illustrated manuscripts. Most of these manuscripts are of French or Flemish origin, but they furnish reliable evidence on English interiors because the governing classes, who were practically the sole possessors of proper furniture, copied the domestic habits of the Continent. English oak was the chief material, but softer woods also were used. A certain amount of furniture was imported from abroad, providing new ideas for the English carpenter and joiner. The furniture usually found in important houses consisted of beds, chests, cupboards, tables, benches, and stools.

The Renaissance. *Italy.* From the beginning of the Renaissance in the early 15th century, there were changes in furniture forms that were to spread over Europe. The growth of a wealthy and powerful bourgeoisie caused the building of more substantial houses and a demand for good furniture. Italian Renaissance furniture shows a strong architectural bias, and the purpose of the piece, as in Roman furniture, was subordinate to its form. The furniture of the early Italian Renaissance is often restrained, with beautiful, simple designs carved in walnut (Figure 66). For more elaborate work, sculpture in low relief and stucco modelled in intricate patterns were much used. The stucco was usually gilded all over and picked out in bright colours.

The cassone, or marriage coffer (hope chest), was a form on which the craftsman's skill was lavished. In addition to elaborate relief work and gilding, these coffers often were painted on the front and sides and occasionally inside the lid as well, with appropriate biblical or mythological scenes. Motifs popular with the Italian carver included cupids, grotesque masks, scrolled foliage, and strapwork. The fixed writing desk is the forerunner of the writing bureau, which became an indispensable article of furniture as writing became more general.

Renaissance chairs

A type of chair called a sgabello was much favoured at this time in Italy. The seat was a small wooden slab, generally octagonal, supported at front and back by solid boards cut into an ornamental shape; an earlier variety was supported by two legs at the front and one in the rear; a solid piece of wood formed the back (Figure 24). Another chair of the period was the folding X-shaped chair, sometimes called a Dante chair. Tables were generally oblong, supported by columns, consoles (brackets), or terminal figures, with a long central stretcher running from end to end. Italian Renaissance furniture forms reshaped the furniture of the remainder of Europe.

France. The furniture of France was among the first to be influenced by the Italian Renaissance. Louis XII and many of his court visited Italy and soon took Italian artists and craftsmen and works of art into France. The French Renaissance of furniture can be divided into two stages. First was a period of transition and adaptation; during the reign of Louis XII and the first part of the reign of Francis I, the pieces were basically Gothic in form, and Gothic ornament was mixed with the cupids, medallion heads, and grotesque decorations of the incoming Renaissance style. During the second phase, from the end of the reign of Francis I, the new style displaced the Gothic. The more exuberant arabesque shapes of Renaissance decoration, however, gave way to increasingly architectural design, and oak was almost entirely superseded by walnut. Centres of furniture making were established at Fontainebleau, where Francis I employed several Italian artists and craftsmen; in Île-de-France, headed by the work of Jacques du Cerceau; and in Burgundy, where, led by the craftsman and designer Hugues Sambin, design was influenced by the Renaissance style evolved in the Netherlands.

16th-century French furniture

French furniture of the 16th century was remarkably graceful and delicate; it was enriched with inlay of small plaques of figured marble and semiprecious stones, sometimes with inlay or marquetry of ivory, mother-of-pearl, and different coloured woods.

Chairs began to be lighter in design; the back became narrower, the panelled sides and base were replaced by carved and turned arms and supports, and legs were joined by stretchers at their base. A specialized chair known as a caquetoire, or conversation chair, supposedly designed for ladies to sit and gossip in, had a high, narrow back and curved arms.

Elaborately carved oblong tables were supported by consoles or fluted columns connected by a stretcher surmounted by an arched colonnade. Chests decorated in the new style were still widely used, although frequently replaced by the armoire (a tall cupboard or wardrobe), which was sometimes made in two stages, the upper compartment containing numerous small drawers.

Spain. Because of the long occupation of Spain by the Moors, a style called Mudéjar evolved. While furniture in

SCALA—Art Resource



Figure 66: Renaissance furniture from the Sala dei Pappagalli in the Palazzo Davanzati, Florence, 15th and 16th centuries.

The vargueno cabinet

this style remained in form essentially European, decoration had an oriental flavour. A type of cabinet known as vargueno was typically Spanish. The upper part, in chest form, with drawers inside, had a fall front (a hinged writing surface that opened by falling forward), often elaborately mounted in wrought iron and backed by velvet, with a massive iron lock. The cabinets were richly carved, painted, gilded, and inlaid with ivory in a Moorish manner. There was a tendency for Italian models to be followed in the furniture of the 16th and 17th centuries.

Low Countries. In the 16th century, Italian Renaissance ornament was adopted and transformed by artists and designers of northern Europe, particularly in northern Germany and the Low Countries, who created an independent style of decoration. Strapwork, cartouches, and grotesque masks are characteristic features of this northern Renaissance style, and are found repeatedly in the pattern books of German and Flemish artists of the time—books of ornament which circulated among and influenced metalworkers, carvers, plasterers and furniture makers throughout the north.

Heavy oak tables, sometimes draw (extension) tables, had massive legs and solid stretchers. Beds were heavily draped to provide privacy, as the bed might be located in any room of the house. Folding wooden chairs and low stools, with more or less elaborate turnery, were still used, besides a new type with baluster-formed or twisted legs and arms, and straight backs heightening through the 17th century.

England. The Italian Renaissance did not affect the design or ornament of furniture in England until about 1520. Evolution from the Gothic style was a gradual process, influence coming first from Italy and, in the second half of the 16th century, from the Low Countries. In the early stages, furniture remained Gothic in form, though Italian motifs slowly replaced the older Gothic ornament. Many pieces of early Renaissance English furniture combined linenfold panelling with medallion heads and Italianate cupids, but by the middle of the century both new ornament and new forms had replaced the medieval style. About the middle of the century the direct influence of Italy weakened, and its place was taken by that of the Low Countries. The northern style of Renaissance ornament was propagated in England by pattern books, immigrant workmen, and imported Flemish and German furniture, and before long it was adapted by English craftsmen into an individual and peculiarly English style.

Styles of the reign of Elizabeth I

Characteristic of this style is the enrichment of every surface with flamboyant carved, turned, inlaid, and painted decoration, which strongly reflects the spirit of the English Renaissance. During Elizabeth I's reign there was a considerable and fairly widespread increase in domestic comfort, to be seen in improved construction, multiplication of types, and the tentative beginnings of upholstered furniture. A series of inlaid chests with perspective architectural scenes, often called nonesuch chests, were either imported from Germany or made by German workmen in England. They were influential in propagating the technique of inlaid decoration, which by the end of the century was being applied to every type of furniture.

Apart from the gradual change from Gothic to Renaissance ornament, the 16th century produced several changes in the design and construction of individual types. Chairs became slightly more common, though even in Elizabeth's own palaces, stools were the usual form of seating. From the box chair evolved a type in which the arms and legs were no longer filled in with panelling but which had plain or turned legs, with shaped arms resting on carved or turned supports. The backs of chairs were still panelled and decorated with carving and inlay or surmounted with a wide and richly carved cresting. Folding chairs, X-shaped and of varying construction, were also used. Chairs without arms, called farthingale chairs, were introduced in the early 17th century to accommodate the wide skirts, called farthingales, that were popular at the time. Farthingale chairs had upholstered seats and a low, rectangular upholstered back raised on short supports a little above the seat. Armchairs of similar design were made. Turkey work (a type of needlework) and velvet were usually employed for upholstery.

Early in the 16th century a new style of bed design appeared; the greater part of the frame was left exposed and was enriched with carving and other decoration, making the frame itself an important part of the design. Favourite carvers' motifs for beds and other types of furniture included strapwork, grotesque masks, and caryatids (draped female figures), bulbous turned pillars and supports, arcading (decorating consisting of arches or arcades), and patterns of scrolled foliage. The heavily turned "cup and cover" motif is frequently found on bedposts in the later 16th century. The cumbersome Gothic trestle tables were replaced by "joyned tables," with tops fixed to the frames. Draw tables, which could be conveniently lengthened by pulling out the two leaves concealed under the top, were also introduced. Table legs and sides were decorated with carving and inlay, and the cup and cover motif is often found on the legs. Various types of cupboards were made, usually in two stages, or levels. In court cupboards both stages were left open. A simple form of chest of drawers was introduced about 1620.

17th century: the Baroque style. During the 17th century, the Baroque style had a marked effect upon furniture design throughout western Europe. Large wardrobes, cupboards, and cabinets had twisted columns, broken pediments, and heavy moldings. In Baroque furniture the details are related to the whole; instead of a framework of unrelated surfaces, each detail contributes to the harmonious movement of the overall design. The Baroque style was adopted in the Low Countries in the 1620s and extended late into the 17th century, when Germany and England began to develop it. It owed much to the Oriental influence that swept over Europe in the 17th century, when several maritime countries, particularly Portugal, Holland, and England, established regular trading relations with India and the Far East. Lacquered furniture and domestic goods were imported from the East, where Oriental craftsmen also worked in a pseudo-European style from designs supplied by the traders. Before the end of the 17th century, Oriental decorative techniques were being widely imitated in Europe, and the roots of the "Chinese taste" were firmly entrenched. Heavy tropical woods were also brought to Europe, and from these, furniture was made that borrowed much from the prevailing taste for Oriental elaboration.

Influence of Oriental decorative techniques

Flanders and Holland. The early Flemish Baroque fur-

Foto Ferruzzi



Figure 67: Flamboyantly carved late Baroque chair made of boxwood by Andrea Brustolon, Venice, *c.* 1690. In the Ca' Rezzonico, Venice.

niture, dating from the second quarter of the 17th century, was but a slight adaptation of the late Renaissance style. Typical are the oak cupboards with four doors and the chairs with seats and backs of velvet or leather held in place by nails.

In Holland the Baroque style did not encroach on late Renaissance furniture until nearly 1640. Dutch furniture of this period can be distinguished by its simpler design and a preference for molded panels over carved ornament. Later, marquetry decoration and walnut-veneer surfaces became the most common decorative treatments. At the end of the century lacquered furniture became popular.

Italy. Though it was in Italian architecture, painting, and sculpture that the Baroque style was evolved, Italy was not the first to apply this style to furniture. But by the mid-17th century Italy was producing flamboyantly carved, painted, and gilded furniture, decorated with such typical motifs as cupids, acanthus, shells, and boldly drawn scrolls (Figure 67), and was further enriching chairs and stools with fine-cut velvets and table tops with marble or *pietra dura* (a mosaic-like technique in which coloured stones are cut and shaped and inlaid in a design). Chairs and stools with exaggerated scrolled arms and legs, and handsome walnut and ebony cabinets and cupboards with carved decoration on the pediments, friezes, and corners and sometimes inlaid with marble or *pietra dura* set in molded panels, typify the Italian furniture of the later Baroque phase (Figure 68).

France. In France the Italian influence of the 16th century was gradually assimilated, and a national style of furniture was evolved that soon spread its influence into neighbouring countries. The reign of Louis XIII, covering most of the first half of the 17th century, was a time of transition. The Gobelins factory was founded by Louis XIV for the production of deluxe furniture and furnishings for the royal palaces and the national buildings. The painter Charles Le Brun was appointed the director in 1663. Furniture was veneered with tortoiseshell or foreign woods, inlaid with brass, pewter and ivory, or heavily gilded all over. At times it was even completely overlaid with repoussé (formed in relief) silver. The name of André-Charles Boulle is particularly associated with this style of decoration. His cabinets and tables were completely covered by sheets of tortoiseshell and brass cut into intricate patterns so as to fit into one another, the tortoiseshell alternately forming the pattern and the ground; hence the two types, boulle (buhl) and counterboulle. The light, fanciful designs of the architect and designer Jean Berain were much used for this work. Heavy gilt bronze mounts protected the corners and other parts from friction and rough handling, and provided further ornament.

Gobelins factory

Turners (Photography) Ltd.



Figure 68: One of a pair of Baroque cabinets inlaid with *pietra dura* made for Louis XIV by the Italian furniture maker and sculptor, Domenico Cucci, at the Gobelins factory, France, 1681–83. In the collection of the Duke of Northumberland, Alnwick Castle, Northumberland, England.

England. After the Restoration, from 1660 onward, there was almost revolutionary progress in English cabinetmaking, as it came to be called at about this time. On its return, the exiled court introduced French and Dutch fashions, and the English craftsmen were considerably helped in supplying the tastes of the nobility by a large influx of foreign workmen. Furniture became lighter, more highly finished, and better adapted to varying needs. The general increase in technical skill of the cabinetmaker between 1660 and about 1690 is astonishing. Walnut was the favourite wood, though the use of oak continued in the country districts for many generations. New processes appeared, notably veneering wide surfaces with thin sheets of wood into which floral patterns in marquetry often were inserted. In the earlier period of the Restoration these patterns were large, but toward the end of the century they grew smaller and more intricate, leading eventually to the type of marquetry made up of numerous small scrolls and called seaweed marquetry.

The passion for colour found an outlet in lacquer decoration in England as in other European countries. The importation of works of art from the East had begun in Tudor times but was of little account until after the Restoration, when the taste became widespread. The diarist John Evelyn and others reported their friends' houses to be furnished with Indian screens or panelled in the finest "japan" (the process that imitated Oriental lacquery was called "japanning" in England).

New forms of furniture

New forms of furniture began to develop: the daybed, a form of couch with an adjustable end; the winged armchairs; the upholstered armchair called in the 17th century a sleeping chair; and, a little later, toward the end of the century, sofas with backs and arms carried comfort a step further. Velvet, silks, and needlework were the usual materials for upholstery. Various kinds of writing furniture were rapidly developed, including toward the end of the century, the bureau with enclosed desk and interior fittings of small drawers and pigeonholes.

Chests of drawers came into more general use. Mirrors were no longer rarities, though glass remained expensive. The frames were carved, lacquered, or decorated with marquetry. Fashions succeeded each other with great rapidity. Chairs show these changes most clearly, developing in a brief period from mere seats of Charles II, while, later, straight tapering baluster forms were used. In the grander beds of this period, the tester (canopy), back, and posts were covered with material. The beds were of enormous height with elaborately molded cornices and had ostrich plumes or vase-shaped finials at the corners of the tester. These state beds were strongly influenced by the designs of the French architect Daniel Marot, who went from France to England to work for William and Mary.

During the late 17th century and on into the first half of the 18th century, a certain amount of elaborately carved and gilded furniture, much influenced by the style of Louis XIV, was produced in England (Figure 69). Foremost among the makers of this deluxe furniture were three cabinetmakers: John Pelletier, Gerrit Jensen, and James Moore. Toward the end of the 17th century, during the reign of William and Mary, Baroque furniture tended to become simpler and the use of ornament was somewhat restrained. At the beginning of the 18th century, during

Queen Anne style



Figure 69: Late Stuart style dining room, Belton House (1685–89), near Grantham, Lincolnshire, England.
Edwin Smith

the reign of Queen Anne, a new and simpler style arose, much influenced by the contemporary furniture of Holland. Carving and applied ornament were reduced to a minimum and the beauty of a piece was made to rely on carefully designed curved lines and the colour of fine walnut veneers. The cabriole leg, originally devised in classical times and based on the curve of an animal's leg, was introduced into England from the Continent about 1700. Terminating in a claw-and-ball or paw foot and soon discarding the stretcher, it was widely used on chairs and tables and for every kind of support. The stretcher had become obsolete because of improved joining and gluing. Chairs had hooped uprights, and fiddle-shaped splate curved to support the back. Tallboys, or double chests of drawers, cabinets fitted with shelves, and bureaus in two stages met the demand for greater convenience, as did a new range of dining, card, and other tables.

The American colonies. As in all colonial settlements, the furniture of the American colonies reflected the style preferences of the individual national groups. This influence, coupled with the existence of new materials and the time lag in transmitting styles and tastes from the home country, in some instances produced highly individual furniture.

Information in inventories and wills about 17th-century furniture of the English colonies indicates that it existed in its simplest forms—stools, benches, tables, cupboards, and a few chairs (Figure 38). This furniture, often made of oak, recalled the tradition of Elizabethan England and was turned and decorated with chip carving, often picked-out in earth colours. By the end of the century, pine, maple, and other woods were used.

The Dutch and Scandinavian settlers carried with them individual furniture forms whose influence remained local.

By 1700 the effect of French and Dutch fashions on late Stuart furniture in England had become evident in the American colonies. Fashion consciousness appeared, though for decades to come the furniture of the average colonial home kept to the earlier tradition evolved from medieval joining. The box chest was succeeded by the chest of drawers, often placed on a stand with turned legs. Chairs began to replace stools; and the early heavy, turned, and wainscot (panelled back) types gave way to simplified versions of the high-back scrolled forms of the English Restoration fashion. The daybed appeared with its upholstered pad. Small folding tables, cabinets, and the tiered dresser to store and display tableware testify to the rapidly increasing standard of comfort among the more prosperous. Carved surface decoration was largely replaced by colour, through the use of paint, veneers, or inlays of contrasting wood.

Awakening of fashion consciousness

These innovations accompanied the use of the cabriole, or reverse curve, which, about 1725, became the favoured form for legs of chairs, tables, cabinets, and stands. At first it had little or no carving and a simple paw foot, but the design was elaborated, and this cabriole leg became the principal feature of the so-called Queen Anne style that dominated colonial furniture designs until the Revolution. Walnut became the principal wood of the early 18th century.

Influence of French furniture

18th century: the Rococo style. The influence of French furniture was predominant in Europe during the 18th century. In the second half of the century England played a leading role in establishing the Neoclassical style, and for supreme craftsmanship provided an inspiration to workshops in several countries; but in the diffusion of the two styles, the Rococo and the Neoclassical, French designs were universally imitated, with varying degrees of success.

France. The transitional phase in French furniture from Baroque to Rococo is called Régence. The heavy, monumental style of the earlier part of Louis's reign was gradually replaced by a lighter and more fluent curvilinear style. The leading exponent of the Régence style was Charles Cressent, *ébéniste* ("cabinetmaker") to the regent Philippe II, duc d'Orléans. In his work the ormolu (a brass imitation of gold) mounts, so important a part of the design of French furniture in the 18th century, became equal to if not more important than, the marquetry decoration of the carcass. The curvilinear form was introduced not only to externals, such as legs and supports, but, in the bombé (rounded sides and front) commodes that first appeared during this period, to the case itself. High-quality marquetry in coloured woods replaced ebony.

Characteristics of the Rococo style

The Rococo style, a development of the Régence, affected French furniture design from about 1735 to 1765. The word is derived from rocailles, used to designate the artificial grottoes and fantastic arrangements of rocks in the garden of Versailles; the shell was one of the basic forms of Rococo ornament. The style was based on asymmetrical design, light and full of movement. The furniture of this period was designed on sinuous and complicated lines. Designs of Juste Meissonier, goldsmith to Louis XV, sculptor and architect, were instrumental in creating the Rococo. The repertory of ornament was large and included the C-scroll, scrolled foliage, floral motifs, ribbon, and, on occasion, trophies formed of musical instruments or gardening implements. The Rococo Chinese taste had conventions of its own: pagodas, exotic birds, Chinese figures, icicles, and dripping water. The graceful bombé commode, often with marble top and two or three drawers, the surface enriched with finely modelled ormolu mounts, was popular. Under Cressent's influence the mounts predominated, though later in the century the marquetry decoration gained first importance. Commodes and other pieces were decorated with marquetry of floral or geometrical patterns, or sometimes with lacquer decoration, again combined with ormolu mounts. The most celebrated makers of mounts during Louis XV's reign were Jacques Caffieri and his son Philippe. Jean-François Oeben was made *ébéniste du roi* (cabinetmaker to the king) in 1754; a pupil of Boulle, he was the most celebrated cabinetmaker of the period (Figure 70).

England. About 1720, mahogany was imported into England and slowly superseded walnut as the fashionable wood for furniture. The Palladian (after the Italian Renaissance architect Andrea Palladio) interiors demanded

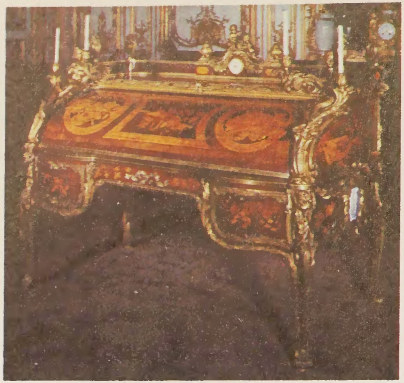


Figure 70: Rococo writing desk, the *bureau du roi*, with intricate pictorial marquetry and elaborate ormolu mounts made for Louis XV, begun by Jean-François Oeben in 1760 and completed by Jean-Henri Riesner, 1769. In Versailles.
Giraudon

furniture more striking and larger in scale than the walnut-veneered pieces of the early 18th century. Inspired by the interiors of French and Italian palaces, architects such as William Kent began to design furniture. The design was classical, in keeping with the traditions of Palladio and the English architect Inigo Jones; the ornament was Baroque. At Holkham Hall in Norfolk, Rousham Hall in Oxfordshire, and elsewhere, Kent's furniture may be seen in its proper environment: gilt mirrors and side tables with sets of chairs and settees covered with patterned velvets matching the grandeur of elaborate architectural Palladian interior decoration.

Despite the resistance of the Palladian classicists who deplored its asymmetrical principles, in the 1740s the Rococo style crept into English decoration and furniture design. During this decade pattern books of ornament in the full Rococo style by Matthias Lock and Henry Copland were published in London; and in 1754 Thomas Chippendale published his *Gentleman and Cabinet Maker's Director*, which provided patterns for a wide range of English furniture in the Rococo style and its Chinese and Gothic offshoots. During the following years several similar works were published by such craftsmen and designers as William Ince and Thomas Mayhew, Thomas Johnson, and Robert Manwaring. The Rococo style was firmly established in England throughout the 1750s and into the 1760s. Chippendale and other cabinetmakers borrowed not only ornament from the French rocaille but designs for individual types. Chippendale's fame rests largely on his publication, though in fact it has now been more or less conclusively proved that he himself was not responsible for the designs, but employed two other designers, Lock and Copland. There were several cabinetmakers—for example, William Vile and John Cobb—whose only memorial is a small quantity of furniture attributable to them. Though it has become the practice to speak of a Chippendale chair or a Vile commode, this does not imply that the pieces were actually made by these craftsmen but that they were made in their workshops.

The work of Thomas Chippendale

By mid-18th century every act of the day that necessitated the use of furniture was catered to by some specialized piece, while the basic furniture such as chairs, cupboards, beds, and tables were designed and decorated in innumerable forms. The number of variants on the Rococo chair splat runs into several hundreds. The ingenuity of the cabinetmaker and carver knew few limitations.

An offshoot of the Rococo style, the Gothic taste was particularly well developed in England. Starting early in the century as a literary device, in the 1740s it began to take more solid shape in architecture, interior decoration, and furniture. As with furniture in the Chinese taste, Gothic furniture bore no relation to its medieval equivalents; the ornaments, such as tracery and cusped (a point formed by the intersection of two arcs or foils) arches, applied to furniture were borrowed from Gothic architecture. The Gothic taste was much publicized by the writer Horace Walpole's celebrated villa, Strawberry Hill, in Middlesex, England. Chippendale included designs for furniture in the Gothic taste in all three editions of his *Director*.

The Gothic taste

The American colonies. Shortly after 1750 the earlier cabriole style was transformed by two factors. One was the rapidly increasing popularity of mahogany. The other was the influence of the English version of free Rococo ornament, as reflected in the publication of Chippendale's book of patterns.

While the Southern planter still depended largely upon London for his fine furnishings, the merchants of Philadelphia, New York, Newport, and Boston were well rewarded by their patronage of local craftsmen. In Philadelphia a local version of the Chippendale style was brought to the highest mastery by such craftsmen as Thomas Affleck, Jonathan Gostelowe, Benjamin Randolph, and William Savery. In Newport, Rhode Island, the genius of the John Goddard and John Townsend cabinetmaking families evolved an equally distinctive style by developing a block front decorated by the patterns of the wood grains instead of carving, as used by their contemporaries in Philadelphia. In spite of the Philadelphians' evident desire to match the works of the best London shops, they actually created their own style as distinct from that of England as the innovations of their Newport colleagues (Figure 39). The cabinetmakers of Boston, New York, and the Connecticut valley also produced work of high quality and a definitely local flavour. Maintaining its hold on popular taste until well after the Revolution, this colonial Chippendale retained more of the sturdy elegance of the earlier cabriole style than did its English equivalent. The tendency of English design to massiveness and surface decoration contrasts with the vertical and linear tendency in much colonial design.

Colonial Chippendale

18th century: the Neoclassical style. *France.* The Neoclassical style, sometimes called Louis Seize, or Louis XVI, began in the 1750s. Tiring of the Rococo style, craftsmen of the 18th century turned for inspiration to classical art. The movement was stimulated by archaeological discoveries, by travel in Italy, Greece, and the Near East, and by the publication, all over Europe, of works on the classical monuments. The Neoclassical style, based on straight lines and rectilinear forms and using a selection of classical ornaments, was first applied to French furniture during the 1760s. Classical motifs at first were sparingly applied to furniture of unchanged form, but slowly the curved line of rococo was replaced by a simpler and more severe rectilinear design: chair legs became straight, tapered, and fluted; commodes and other storage furniture were no longer of bombé form. Marquetry was still widely used for decoration, and some cabinets were made of ebony inset with panels of Japanese lacquer. Boulle, which had not been employed in Louis XV's reign, returned to fashion. A greater number of pieces were signed during this period (signing had been made compulsory in Paris), and Jean-Henri Riesener, Martin Carlin, and Jean Saunier were a few of the leading cabinetmakers. Several German craftsmen migrated to France because of the royal patronage, among them Abraham and David Roentgen, Adam Weisweiler, and Guillaume Beneman.

These craftsmen were often directly under the patronage of the king, having their workshops in the cellars of the Louvre. Within the shop there was a division of labour, with one craftsman specializing in furniture construction, another in lacquering, and so forth. The craftsmen and the shop were licensed by the government.

England. The classical reaction, which set in shortly after 1760, reimposed a classical discipline on design, though of a lighter and more delicate touch than that of the previ-

ous classicists, the Palladians. Robert Adam, whose name is inseparably associated with this movement, had, like earlier architects, studied in Italy. There he sought inspiration in the monuments of both classical times and the Renaissance. When given a free hand, he included interior decoration and furniture in his architectural schemes, one of the best examples being his alterations and redecorations at Osterley, Middlesex, where he provided harmonious designs for even the lock plates and chimney pieces. His furniture makes restrained use of classical ornament; but paterae (disks with a design in relief or intaglio), husks (a drop ornament made of whorls of conventionalized foliage usually in a diminishing series), rams' heads, and urns are less eloquent of the change than the symmetrical structural lines. Marquetry, ormolu mounts, and painting were employed as decoration (Figures 37, 71). Adam's furniture was copied and modified by contemporary cabinetmakers such as George Hepplewhite in his *Cabinet-Maker and Upholsterer's Guide* (1788).

In the last 20 years of the 18th century there was a tendency toward greater refinement, lightness, and delicacy in furniture design. Symmetry of form and excellence of proportion were retained for the most part. Heart- and shield-shaped backs on chairs and settees and tapered and fluted supports for tables and other pieces are characteristic; feathers, wheat ears, and shells are prominent in the painted or inlaid decoration. This refinement, strongly feminine in character, is represented in Thomas Sheraton's *Cabinet-Maker and Upholsterers' Drawing Book* (1791).

The work of Thomas Sheraton

The United States. The new classicism of Robert and James Adam came into vogue in the new republic during the last years of the century (Figure 43). The shipowners and merchants of Salem, Boston, and New York equipped their mansions with the work of Samuel McIntire, John Seymour, and Duncan Phyfe, each of whom produced individual interpretations of the Hepplewhite-Sheraton mode. This early Federal style is characterized by small-scale rectangular design and by a preference for light-toned wood finishes. Surfaces are generally unbroken but decorated with bandings and inlays of contrasting woods, or in Phyfe's case with low relief carvings in the Adam manner. The most typical pieces are the sideboard (a piece of dining room furniture with compartments and shelves for dishes) and the small secretary desk, both of which developed a peculiarly American form.

19th century. The Empire style began in Paris about the time of the Revolution and quickly spread throughout Europe, each country adapting it to its own national taste. In England it is commonly called the Regency style. Two French architects, Charles Percier and Pierre Fontaine, who designed the furnishings for the staterooms of Napoleon, contributed in great measure to the creation of the style. Their ideas were incorporated and propagated in *Recueil de décorations intérieures* (1801 and 1812; "Collection of Interior Decoration").

Basically the new style was a continuation of the Neoclassical style, with a much stronger archaeological bias, leading to direct copying of classical types of furniture; to this was added a new repertory of Egyptian ornament, stimulated by Napoleon's campaigns in Egypt. Mahogany-veneered furniture with ormolu mounts assumed the shapes of Roman, Greek, and Egyptian chairs and tables, with winged-lion supports and pilasters headed with sphinxes' busts or palm leaves; where no classical prototypes existed, contemporary designs were enlivened with classical ornament.

Characteristics of the Empire style

In England, Thomas Hope, an amateur designer with some knowledge of antiquities, was the chief exponent of the Regency style and entirely decorated his country house, Deepdene, Surrey, in it. When the fashion was taken up by cabinetmakers, the results were often woefully incongruous. Mahogany and rosewood were used with bronzed or gilt ornament, and metal inlay, a cheaper technique, replaced inlay and marquetry. Along with this style came a renewed enthusiasm for the Chinese taste, as best exemplified in the furniture and decoration of the Brighton Pavilion (Figure 41). In the final stages of the Regency style, both the design and construction of furniture in England and on the Continent showed signs of heaviness and overelaboration that heralded the general decline throughout Europe in the 19th century.

In the United States the style was widely adopted. Its chief native practitioner was the New York cabinetmaker Duncan Phyfe, who in the first decade of the century produced furniture for the wealthy of his city. His designs gave a unique interpretation to Empire ideas (Figure 72). French cabinetmakers, such as Charles-Honoré Lannuier, emigrated to the United States at this time and produced furniture in a stricter French style.

The work of Duncan Phyfe

By the 19th century, with increases in the efficiency of transportation and communication, styles became more

By courtesy of the Metropolitan Museum of Art, New York, Rogers Fund, 1932



Figure 71: Neoclassical dining room from Lansdowne House, Berkeley Square, London, designed by Robert Adam, *c.* 1765. In the Metropolitan Museum of Art, New York.



Figure 72: Parlor furniture made by Duncan Phyfe for Samuel A. Foot, New York, 1837. In the Metropolitan Museum of Art, New York, 19th Century Centennial Exhibition, 1970.
By courtesy of the Metropolitan Museum of Art, New York, 1966 Purchase, L.E. Katzenbach Foundation Gift

universal in their adoption but still maintained national and regional differences.

The Empire style, which carried over into the 19th century, began a series of styles that revived form and decoration from the past. This reinterpretation often resulted in a product removed from the principles of the original style. The introduction of the machine and of the factory method sometimes brought about a decline in quality in furniture production.

Biedermeier style

The Biedermeier style, which originated in Germany and Austria, flourished in the prosperous middle class homes of Europe from about 1815 to 1848. This style is characterized by classical simplicity. Chairs had curved legs, and sofas had rolled arms and generous upholstery. Mahogany veneers and light birch, grained ash, pear, and cherry were used. The design and much of the ornament were influenced by the Empire style, in particular the Grecian element. The style took its name from "Papa Biedermeier," a fictitious character whose column, offering opinions on taste in furniture, appeared in Austrian newspapers.

In the 1820s there was a revival of the Gothic style, which in England was partly stimulated by romantic literature such as the novels of Sir Walter Scott. Losing all the lightness and humour of the mid-18th-century Gothic revival, heavy medieval motifs were profusely and indiscriminately applied to every type of furniture.

A series of other revival styles followed the Gothic. The Rococo revival was one of the most popular; it borrowed the curvilinear elements of the French Louis XV style, especially the cabriole leg, and restated them in a heavier idiom. Entire suites of this furniture were fashioned in mahogany, rosewood, and walnut, the price being highly dependent upon the amount of carving on the frame.

During the first half of the 19th century (the exact date is unknown), metal springs were introduced into furniture construction. The spring construction made chairs and sofas much more comfortable than had the stuffing employed by cabinetmakers during the 18th century.

Another technical improvement introduced into furniture design was the use of plywood. Plywood had great strength and stability and could be more intricately curved than a natural piece of wood. One of the chief exponents of this technique in the United States was John Henry Belter, who was born in Germany in 1804 and served his cabinetmaker's apprenticeship in Württemberg. He reached a height of popularity in the 1850s. Belter's work is mainly in the Louis XV revival style.

Michael Thonet, an Austrian craftsman, experimented with bending layers of veneer in Boppard, Germany. Thonet was successful in perfecting a process for bending solid beechwood by heat into curvilinear shapes. His chairs, popular during the latter half of the 19th century, are still made.

Elizabethan and Louis XIV revival furniture was also very popular. The Baroque twisted upright was one of the chief elements employed. The straight, turned leg was also reintroduced. This elaborately upholstered furniture was produced in suites and was blocky and square in its overall form, in contrast to the Rococo revival form.

The Louis XVI style was reintroduced in suites of furniture with round tapering legs, oval backs on chairs and sofas, and elaborate upholstery. The Louis XVI leg was often used on comfortable upholstered furniture whose structure consisted primarily of a flexible metal, or "Turkish," frame. The only wood visible on this furniture was in the legs, the remainder of the frame being completely upholstered. In such furniture the art of the upholsterer reached its height through the use of elaborate tufting, tassels, and braids.

The influence of William Morris

The English poet and artist William Morris has been called the father of the modern movement. Critical of the shoddiness of the machine-produced goods of his own day, he turned for inspiration to the handcraftsmanship of the Middle Ages and, basing his own work on their designs and methods, attempted to revive a respect for fine craftsmanship and to stir the aesthetic sense of his contemporaries (Figure 42). His influence, though important, might have been greater if, instead of turning away from the machine, he had applied his high ideals to discovering a way in which machines might be used to the best advantage. Morris' followers in the field of cabinetmaking included such designer-craftsmen as Ernest Gimson and the Barnsley family who, working with a few assistants, produced small quantities of high-quality handmade furniture, the craftsmanship of which has never been rivalled. The example of Morris and his followers was so widely copied on the Continent that many people believe modern furniture design originated exclusively there.

During the third quarter of the century, there was a movement in England toward greater simplicity and aesthetic beauty in furniture. The straight and simple lines of Japanese design served as a source of inspiration. The result was the aesthetic, or artistic, style; its chief exponents, producing both designs and furniture, were Edward Godwin and Christopher Dresser.

Art Nouveau style

Henry van de Velde, a Belgian architect and designer, followed in the footsteps of William Morris and was the conscious propagandist of the Art Nouveau style, which flourished from about 1893 to 1910. Characterized by moving, sinuous curves, the style found its inspiration in organic and natural forms and in the Japanese prints that were so popular in Europe during the third quarter of the 19th century. Van de Velde's furniture was often designed *en suite* so that it would give an effect of totality to a room. The interiors of a house in Brussels, created by another Belgian architect, Victor Horta, well illustrate the sinuous curves and natural forms employed by the Art Nouveau designers. The movement was also adopted in France where Hector Guimard was one of its chief exponents. A variant of the style is seen in furniture produced by the Scottish architect Charles Rennie Mackintosh (Figure 73). The Art Nouveau style in furniture design was not as popular in England or in the United States as it was on the Continent. (J.T.B.)

By courtesy of the University Art Collections, University of Glasgow, Scotland

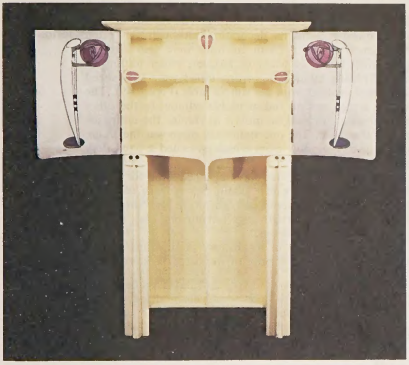


Figure 73: Art Nouveau painted oak cabinet with coloured glass by Charles Rennie Mackintosh, 1902. In the University Art Collections, University of Glasgow.

Modern. After the late 19th century, furniture design in the West was divided into two main categories: revivals of past styles—only occasionally precise reproductions, more often free adaptations; and various expressions of changing modern life. The latter category absorbed the best as well as the most progressive talents of the era.

Modern furniture design after World War I was of three kinds: functionalist modern—progressive, adhering to an aesthetic of the machine and often designed by leading architects; transitional modern, which came to be called contemporary and was infused with elements from the past; and commercial modern, called "Borax" because hawkers of that cleanser used to offer premiums, and the word became associated with extra values which commercial furniture often offered by the manner in which it was advertised, or in overblown forms and gaudy veneers. All furniture design was influenced by the social and economic trends of the era: formal living declined; mechanization of household labour expanded; living spaces shrank, particularly in height; and home entertainment became important. After World War II, especially, people married at a younger age, total population growth accelerated, and a generally rising standard of living was enjoyed by a vastly enlarged middle-income group. Furniture became smaller, lighter, easier to maintain, and more widely distributed.

Functionalist modern. About 1925, a new rationality began in furniture design, stimulated by the emergence of progressive experiments typified in the works and theories of the Bauhaus, a revolutionary German school of arts and crafts established in 1919 and staffed by leading architects, designers, and painters until Hitler closed it in 1933. Bauhaus instruction used crafts as experimental techniques and trained students to design for mass production. Low price levels, maximum utility, good quality, and simple, clear forms were considered essentials of well-designed consumer goods. The celebration of modern technology in progressive design was the most effective accomplishment of the Bauhaus. Forms, colours, and materials hitherto confined to shops and laboratories were introduced into homes and offices with programmatic earnestness and considerable stylishness. Tubular chrome-plated metal, black Bakelite, and large unframed planes of glass were typical. Much furniture used at midcentury in reception rooms, terraces, kitchens, or dining alcoves derived from Bauhaus originals. The availability of wood in Scandinavia led, in the 1930s, to similar rational, modern furniture, using a variety of laminating techniques. Related, more ambitious experiments in three-dimensional molding of wood laminates were undertaken in the United States around 1940. Then wartime austerity enforced a salutary simplicity.

The Bauhaus designs

After World War II, earlier design activity resumed. Scandinavian designers abandoned advanced technology for a time and launched a victorious campaign for sculptured, solid-wood furniture in matte finishes that notably enlarged the vocabulary of progressive design. Italian furniture was similar in trend, more open to structural and technological experiments but more accented and less acceptable generally. American modern furniture achieved its first international influence in molded plywood and plastic chairs and in semiarchitectural storage units.

Scandinavian designers

Functionalist modern furniture consciously related itself to progressive architecture, which aided its steady growth in the third, fourth, and fifth decades of the 20th century; at the same time it was also encouraged by friendly periodicals, shops, and museums. Educational and cultural agencies earlier in the century had generally opposed modern design, but gradually there was a change in attitude and by mid-20th century, it was accepted.

Transitional modern. Conservative in style (but not imitative), well-constructed, and carefully finished, the best modern furniture earned its reputation of being in good, correct taste. Often relying on handcraft details and on wood, most factories used speeded-up variations of earlier cabinetmaking operations. This, along with the United States' emphasis on artificially stimulated obsolescence, affected all modern design between World Wars I and II. As in the case of stylistic revivals, favourite sources of inspiration for transitional modern were late 18th- and early 19th-century court and country house furniture, with variations in Chinese and Rococo. This furniture served a wide public that found the avant-garde forms and materials too cold and "clinical."

Commercial modern. Most modern furniture designed between 1930 and 1940 was bulky, bulbous, glowingly coloured, glossily finished, and rich with hardware or shiny fabric. It pleased the public but not critics and connoisseurs. Gradually, and more noticeably after 1945, stylistic details filtered down from more progressive design levels to appear as commercial fads, such as sectional seating and storage units, spidery metal frames, and plastic-shell seats; the Victorian whatnot (set of open shelves for the display of bric-a-brac) was revived, freestanding and rectilinear, as the room divider. Convertible sofa beds and radio and television cabinets were almost all designed in the commercial manner. The innovation of foam upholstery was bitterly fought by union workmen around 1940, but in 15 years had become commonplace in sleeping and seating furniture.

In time a continual flow of new production methods effected basic changes. Lighter masses, thinner silhouettes, and new forms made possible by new materials as well as new technologies seemed to put modern furniture design

on the threshold of a new era. By 1970, however, faddism and commercial versions of bizarre and bloated shapes in seating furniture again ushered in a new brand of "Borax." (E.J.Wo.)

EASTERN

China. Remarkably little systematic study has been made of Chinese furniture. Its origins remain comparatively obscure, its workshops mostly unrecorded, its designers unknown; consequently, its dating is extremely difficult. Most of the forms of Chinese furniture, such as the low table and the covered bed, are found in the oldest Chinese paintings in existence; the designs have been remarkably conservative throughout the ages.

Two types of Chinese furniture

Chinese furniture can be divided into two main types: lacquered wood pieces either inlaid with mother-of-pearl or elaborately carved, and plain hardwood pieces.

Of the first, almost nothing is known, and dating of pieces is possible only from the designs of decorative motifs, such as dragons and peonies, and from their background motifs. The most important historically in this class are black lacquer pieces inlaid with mother-of-pearl that have been preserved in the Imperial repository (Shōsō-in) in Japan from the 8th century. Of the red lacquers, such as seats and tables, the earliest pieces date from the Ming dynasty (1368–1644); their workmanship is characterized by softer contours and freer, more spirited designs than the later pieces of the Ch'ing dynasty (1644–1911/12) (Figure 74). These lacquered objects influenced European cabinetmakers.

Plain hardwood furniture is frequently encountered. Its deserved popularity both in China and the West has been won by its classical simplicity, reserved ornament, and lack of pretense. In these products of the finest workmanship, purity of line, plastic strength, and a flawless polish produce a harmonious, solid effect (Figure 48).

A Chinese house requires less furniture than a Western house. Correspondingly, the types of furniture are fewer, being limited mainly to wardrobes, chests, tables both high and low of all types and shapes (altar and couch tables, for example), stools, beds (sometimes testered with curtains), screens and stools for use by the bed, and chairs.

Although the fundamentals of Chinese joinery must have been formed a millennium before the modern era, the great development in Chinese furniture took place with the introduction of Buddhism from India during the first centuries AD. Before that time the Chinese had sat cross-legged, or knelt on the floor or on stools. Buddhism introduced a more formal kind of sitting on stiff, higher chairs with back rests and with or without side arms. The chests and armoires are superb examples of careful joinery and often have finely worked metal mounts that greatly enhance the beauty of their solid design.

A number of hardwoods were used for the plain furniture: purple sandalwood (the most distinguished); rosewood of many varieties, mostly imported from Indochina and called "old," "new," and "yellow"; redwood; burl (especially for inlay); and so-called chicken-wing wood. Rosewood in its many varieties is perhaps the most frequently encountered and the most popular for its seeming translucence and satin, soft finish. It is above all the faultless workmanship, so typically Chinese, and the fine polish of Chinese furniture that attracts the Westerner. It was the Chinese respect for the spirit of wood and their command of line, curve, and cubic proportions that became the ideal of the 18th-century Western cabinetmaker.

Japan. Japan was one of the few civilizations that did not develop many specialized furniture forms. Instead, the interior architecture of the house, with the garden as its focal point, served the aesthetic and social requirements that furniture has served in many societies. The chief requirement for the few forms that were developed was that they be easily movable.

Japanese furniture forms

Thin mats made of rice straw called tatami covered the floors and were used for sitting. The tatami utilized only natural patterns for decoration, although they often were bound in cloth. Cloth cushions were also used, as were small tables of wood or lacquer, either folding or rigid. Dressing tables and writing tables were specialized forms that evolved from the simple table (Figure 49). The folding screen was an indispensable adjunct to the other furnishings as it could be moved to change the entire aspect of the room. The one stationary piece was the *shoin*, a type of bay window from which extended a fixed desk used for reading.

Japanese furniture forms have changed little for centuries. Because there are few extant pieces from the early periods, information about early furniture is gleaned from literary descriptions, engravings on mirrors, clay images, and graphic representations.

India. India's place in the history of furniture is that of an adapter or transformer of imported Western styles

Photograph, American Heritage Publishing Co., Inc.



Figure 74: The Ch'ien-lung emperor, art collector and skilled painter, with tables, screen, and daybed from his furniture collection, China, 18th century.

rather than a creator of independent styles of its own. Domestic furniture in the sense in which it is known in Europe was not traditional in India before the 16th century, and even such familiar objects as tables and chairs were rarely used until the spread of Portuguese, Dutch, and English furniture.

It was precisely the difficulty of obtaining suitable furniture locally for their settlements that encouraged the European traders to export Western prototypes for copying. It was soon found, however, that the Indian craftsman, although an inaccurate copyist, was a skilled and imaginative adapter of foreign decorative detail. This led to the emergence of an independent Indo-European style of furniture that was much admired for its own sake and subsequently exerted fresh influences in the West. Early Indo-European furniture can be divided into two distinct groups, according to whether the influence was primarily Portuguese or Dutch. (The English did not exert a national influence on styles until the late 18th century.)

Two groups of Indo-European furniture

The Indo-Portuguese group includes a northern Indian, or provincial Mughal, style and a southern, or so-called Goanese, style. The former is artistically the more interesting and includes a variety of furniture decorated with inlaid bone or ivory on ebony and other dark woods. Tables and writing cabinets in the Italian Renaissance form are found in this category because this was the dominant style in Portugal.

The second Indo-Portuguese style, sometimes called Goanese (though in fact more probably made on the Malabar coast, south of Goa), is more stereotyped in form and in decoration. It is distinguished by large and rather cumbersome cabinets of a type known in Portugal as *contador*, the inlay ornament being either geometrical or semiabstract. The Indian contribution to this style is more inhibited and lacks altogether the charm and fancifulness of northern Indo-Portuguese furniture.

Indo-Dutch furniture is easily distinguishable from Indo-Portuguese, since it reflects contemporary Dutch taste as clearly as the latter reflects Portuguese. There are two types of Indo-Dutch furniture. The first, which was made on the Coromandel coast, was mainly in light-coloured woods, the decoration being inlaid bone, incised and lacquered. The second is a style of carved ebony furniture which, although commonly found in India and often thought to be Indian in origin, was in fact made at Batavia (modern Djakarta) in Java, the Dutch administrative headquarters in the East. The carved relief decoration of the ebony furniture is floral in character and closely related to the flowering-tree style of contemporary Indo-Dutch embroidered bedspreads and hangings in which the tulip is prominent.

Indo-Dutch furniture

With the growth of British power in India in the 18th century, all Indo-European furniture styles came increasingly under English influence. Whole suites were made in ivory in the manner of Chippendale and Sheraton, not only for European buyers but also for Indian rulers who increasingly favoured European styles of furniture.

In the 19th century, Indian artistic standards degenerated, as is clearly reflected in the furniture of the period. The emphasis was on decorative elaboration for its own sake and, although much 19th-century Indian wood carving shows great technical skill, this rarely compensates for formlessness and stereotyped ornament.

(J.T.B.)

RUGS AND CARPETS

The word carpet was used until the 19th century for any cover made of a thick material, such as a table cover or wall hanging; since the introduction of machine-made products, it has been used almost exclusively for a floor covering. Both in Great Britain and in the United States the word rug is often used for a partial floor covering as distinguished from carpet, which frequently is tacked down to the floor and usually covers it wall-to-wall. In reference to handmade carpets, however, the names rug and carpet are used interchangeably and are so used in this section, which deals almost exclusively with handmade products. Since such carpets are not always intended for use on the floor, the section extends the term rugs and carpets to cover products intended for other uses as well.

Handmade carpets are works of art as well as functional objects. Indeed, many Oriental carpets have reached such supreme heights of artistic expression that they have always been regarded in the East as objects of exceptional beauty and luxury in the same way as masterpieces of painting have been in the West. Handmade carpets are discussed in this section in terms of their elements of design, material, technique, ornament and imagery, use, and stylistic characteristics in different periods and cultures.

Elements of design

FIELD AND BORDER DESIGNS

Designs usually consist of an inner field—the pattern in the centre of the carpet—and a border. The latter serves, like the cornice on a building or the frame on a picture, to emphasize the limits, isolate the field, and sometimes control the implied movements of the interior pattern. The design of inner field and border must harmonize pleasingly, yet remain distinct.

The border consists of a minimum of three elements: a main band, which varies greatly in width according to the size of the rug and the elaborateness of the field design, and inner and outer guard stripes, decidedly subordinate bands on either side of the main band. The guard stripes may be the same on both sides of the main band or be different. The most common decoration for the field is an allover pattern, a panel composition, or a medallion system. The allover pattern may be of identical repeats (Figure 75), either juxtaposed or evenly spaced, though the latter, while common on textiles, is rare on carpets; or it may be of varied motifs in a unified system (*e.g.*, different plant forms of about the same size), but even this freest type of design almost invariably includes bilaterally balanced repetitions. The varied motif type of design is found most typically in formalized representations of the parks or woods that were a feature of Persian palace grounds.

Types of field decoration

Another type of allover design appears to be entirely free but is actually organized on systems of scrolling stems, notably on the east Persian carpets of the 16th and 17th centuries.

The value of panel subdivisions for controlling patterns had been discovered in a simple rectangular version by the Upper Paleolithic period (*c.* 25,000 BC), and panel systems have been a basic form of design since 4000 BC, when pottery painters were already devising varied systems. On carpets, the lattice provides the simplest division of the field, often a diagonal lattice as on an embroidered

By courtesy of the Victoria and Albert Museum, London

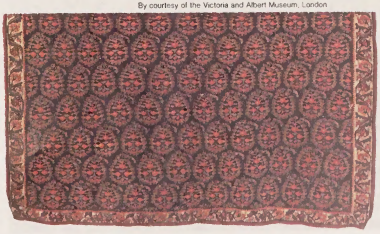


Figure 75: Detail of a Persian *kilim* from Sehna (Sanandaj), Iran, 19th century. A tapestry-woven wool rug with an allover identical repeat pattern of *bâteh* (leaf with curling tip) in rows. In the Victoria and Albert Museum, London. Full size 1.65 × 1.19 m.

carpet found in an excavated tomb (1st century BC to 1st century AD) at Noin Ula in northern Mongolia; the diagonal scheme also appears on Sāsānian capitals and in Coptic tapestries. But a characteristic field design of the Persian court carpets of the Shāh 'Abbās period, the so-called Vase pattern, is constructed from the ogee, a motif that became prominent in Middle Eastern textile design in the 14th century. Simple rectangular panelling—really a large-scale check—is typical of one style of Spanish rugs of the 15th and 16th centuries.

Medallion carpets

The most frequent medallion composition consists of a more or less elaborate medallion superimposed on the centre of a patterned field and often complemented with cornerpieces, which are typical quadrants of the central medallion (Figure 76). But multiple-medallion systems also are developed: either a succession or a chain of medallions on the vertical axis; two or more forms of medallions alternating in bands, a scheme typical of the Turkish (Ushak) carpets of the 16th and 17th centuries; or systematically spotted medallions that may or may not be interconnected or that may interlock so that the scheme logically becomes an elaborate lattice.

Persian carpets of the 15th–17th century commonly have multiple-design schemes; that is, composition systems on two or more "levels." The simplest is the medallion superimposed on an allover design, but more typical are subtler inventions such as two- or three-spiral stem systems, sometimes overlarded with large-scale cloud bands, all intertwining but each carried independently to completion. The finer Vase carpets have double or triple ogival lattices set at different intervals (staggered), each with its own centre, and tangent motifs that also serve other functions in the other systems. What at first sight appears to be a great multiplicity of independent motifs thus proves on careful examination to be ingeniously contrived and firmly controlled.

Occasionally, stripe systems are used, either vertical or diagonal; but this conception is more natural to shuttlewoven fabrics, and, when employed in the freer techniques of rug weaving, it is probably an imitation of textiles.

By courtesy of the National Gallery of Art, Washington, D.C., Widener Collection; photograph: Otto E. Nelson—EB Inc.



Figure 76: Persian silk carpet from Kashan, Iran, late 16th century. The field is decorated with a central medallion, surrounded by a wreath of small cartouches and framed by corresponding cornerpieces. In the National Gallery of Art, Washington, D.C. 2.41 × 1.65 m.

DESIGN EXECUTION

Transferring the design is done in various ways. It can be transferred to the carpet directly from the mind and hand of the craftsman or indirectly from a pattern drawn on paper. Using the latter technique, a rug can be executed directly from the pattern, or the design can be transferred first to a cartoon. The cartoon, or *talim*, is a full-size paper drawing that is squared, each square representing one knot of a particular colour. The weaver places the *talim* behind his loom and translates the design directly onto the carpet. The cartoon is used for reproduction of very intricate designs and as a master pattern for the production of more than one carpet. Many of the finest Oriental rugs, which achieve a magnificent effect through wealth of detail, are thought to have been woven from cartoons drawn by manuscript illuminators. Such methods of transfer result in unavoidable irregularities of pattern that, because they are signs of the artistic individuality of the craftsman, lend a particular charm to the handwoven carpet. The major difference between handmade and machine-made carpets is that the mechanical transfer of design in the latter creates a uniformity of pattern, obliterating signs of individual workmanship.

COLOUR

From earliest times until the late 19th century, only natural dyes were used. Some have come from vegetables such as madder, indigo, sumac, genista, and woad; some from minerals such as ochre; and some from animals, mollusks, and insects. Most have been improved by the addition of various chemicals, such as alum, which fix colours in the fibre. Except for dark brown to black dyes, which have high iron-oxide content that often decomposes fibres, natural dyes have proved to be excellent; they have remarkable beauty and subtlety of colour, and they are durable. Much of the charm of antique carpets lies in the slightly varying hues and shades obtained with these natural dyes. In the 19th century, synthetic, aniline dyes were developed, becoming popular first in Europe and, after 1860, in the East; but their garish colours and poor durability were later thought to outweigh the advantages of brilliance and quick application, and natural dyes regained favour with many craftsmen. More recently, synthetic dyes have been improved.

Natural dyes

Materials and technique

Most carpets are made of sheep's wool, which is durable, dyes readily, and handles easily. Camel or goat wool is rarely used. Too dull to make an attractive pile, cotton's strength and smooth yarn make it an ideal warp (see below); it is used in the East for the entire foundation or for the warp only.

Silk is so expensive that its use is restricted; but no other material produces such luxurious, delicate rugs, displaying subtle colour nuances of particular charm in different lights. Some of the finest 16th- and 17th-century Persian carpets are entirely of silk. It has never been used for knotting in Europe; but often since the 15th century it has augmented wool in the weft of European tapestries.

Linen was used in Egyptian carpets, hemp for the foundations of Indian carpets; and both materials are used in European carpets. Since around 1820, jute has been used in the foundation of machine-made carpets.

Knotted pile carpets, combining beauty, durability, and possibilities for infinite variety, have found greatest favour as floor coverings. Long ago, weavers first began to produce pile fabrics or fabrics with a surface made up of loops of yarn, attempting to combine the advantages of a woven textile with those of animal fleece. Knotted pile is constructed on the loom on a foundation of woven yarns,

of which the horizontal yarns are called weft yarns and the vertical are called warp yarns. Coloured pile yarns, from which the pattern is obtained, are firmly knotted around two warp yarns in such a way that their free ends rise above the woven foundation to form a tufted pile or thick cushion of yarn ends covering one side of the foundation weave. The knots are worked in rows between several interlocking, tautly drawn weft yarns that keep every row of knotted tufts securely in place in the foundation. When a row of knots is tied, it is beaten down against the preceding rows with a heavy mallet-like comb so that on the front the pile completely conceals both warp and weft. When a certain area has been woven, the pile ends are sheared to an even height: short on the more aristocratic type and as much as an inch ($2\frac{1}{2}$ centimetres) on some shaggy nomadic rugs.

Types of knots

There are various ways of knotting the pile yarn around the warp yarn. The Turkish, or Ghiordes, knot is thought to be the oldest. It is used mainly in Asia Minor, the Caucasus, Iran (formerly Persia), and Europe. The Persian, or Sehna, knot is used principally in Iran, India, China, and Egypt. The Spanish knot, used mainly in Spain, differs from the other two types in looping around only one warp yarn. After the 18th century it became extremely rare. The kind of knot used affects the delicacy and tightness of the pile. Knotting each pile yarn by hand is comparable to setting small pebbles in a mosaic, and expert execution is vital in achieving a beautiful finished product. Angular-patterned carpets requiring only a coarsely knotted pile are easier to produce than curvilinear and finely patterned ones, which require finer material and a much more densely knotted pile for clear reproduction of their intricate designs. Some Chinese carpets have fewer than 20 knots per square inch (three per square centimetre); certain Indian ones, more than 2,400. The highest density can be achieved with the Persian, or Sehna, knot.

Brocading can be added to the pile, heightening its colourfulness. The gold and silver thread in this procedure lies flat against the woven foundation, giving the appearance of low relief (Figure 77, left). Metal threads, however, quick-wearing and with diminishing lustre, are less suitable as floor coverings than as hangings.

Many carpets do not have knotted pile. Called *kilims*, they are woven similarly to tapestries. The weft yarns of a given colour area never cross into another area, and if the weft yarns of different colour areas are hooked around adjacent warps rather than around one another or around warp yarn, small slits are created where different colours meet (Figure 77, centre). In fully brocaded Soumak carpets, one or two rows of coloured pattern weft alternate with an invisible functional weft. Brocading with passes of alternate rows giveu a differing direction, or slant, produces a herringbone effect (Figure 77, right).

Embroidery has rarely been used on floor coverings. Embroidered rugs are almost exclusively European and American, except for certain Turkmen *kilims* and Turkish *cicims* (ruglike spreads or hangings). Only relatively strong backings can be used, such as in appliqué work or embroideries done in designs of counted stitching (the cross-stitch, and the gros point and petit point of needlework) that cover the entire surface.

Ornament and imagery

INDIVIDUAL MOTIFS

Three main classes of motifs are used: geometrical; conventional, or stylized; and illustrative, or naturalistic. The geometrical repertoire is built up from variations and combinations of meanders, polygons, crosses, and stars. Meanders, chiefly for borders, range from the simple serration employed from earliest times to fairly complex hooked forms, characteristically the angular "running wave," or

(Centre) The Textile Museum Collection, Washington, D.C., gift of Mrs. David B. Kamick, by courtesy of (left) the Metropolitan Museum of Art, New York, gift of John D. Rockefeller, Jr., 1950, (right) the National Rug and Textile Foundation, Washington, D.C., gift of W. Russell Pickering; photographs, Otto E. Nelson—EB inc.



Figure 77: *Techniques of rug making.*
(Top left) Detail of a Polish carpet, a gold-and-silver brocaded silk rug from Persia, 17th century. In the Metropolitan Museum of Art, New York. Full size 3.96 × 1.77 m. (Bottom left) Enlarged section of above showing the contrast between silken knotted pile and areas brocaded with metal-covered yarns. (Top centre) Detail of a Shirvan wool *kilim* (tapestry-woven carpet), southeastern Caucasus (Azerbaijan S.S.R.), late 19th century. In the Textile Museum, Washington, D.C. Full size 2.59 × 1.47 m. (Bottom centre) Enlarged section of above showing slits produced where two colours meet along the vertical, the yarn of each colour having been returned on the same warp. (Top right) Detail of a wool Soumak carpet, Caucasus, 19th century. In the collection of the National Rug and Textile Foundation, Washington, D.C. Full size 2.29 × 1.75 m. (Bottom right) Enlarged section of above showing the herringbone effect created by the Soumak method of brocading.

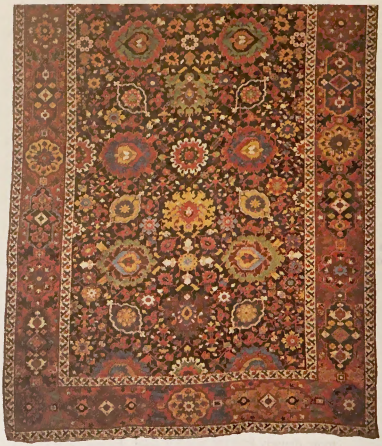
"Greek key," which is also very ancient. Little trefoil (trilobed) motifs are used for guard stripes in the Caucasus, central Iran, and India. Chief among the polygons employed are the lozenge and the octagon. The Maltese cross is frequently used, as is the gamma cross, or swastika. Purely geometrical stars are usually based on the cross or the octagon. Many of these motifs, which are rudimentary and very ancient, may have originated in basket weaving and the related reed-mat plaiting, for they are natural to both techniques; but in rug weaving they have survived chiefly in the work of Central Asia, Asia Minor, and the Caucasus, in both pile-knotted and flat-woven fabrics.

Arabesque carpets

One of the principal stylized motifs in 16th- and 17th-century Persian carpets is the so-called arabesque, an ambiguous term that generally implies an intricate scrolling-vine system. In a common Persian ornamental scheme, two asymmetrical members cross at an acute angle, forming a lilylike blossom, and then describe curves in opposite directions, readily continuing into further scroll systems (Figure 78). This highly individual form was begun in China in the late Chou period (*c.* 600 BC), notably on a few bronze mirrors, and was beautifully developed during the Han dynasty (206 BC–AD 220). It appeared in Persia in the 12th century (on pottery and architectural stucco ornament), possibly influenced by the Chinese form.

Directly traceable to China are the cloud knot and cloud band, or ribbon—both in use by the Han period at least and with a continuous history thereafter. The cloud knot, a feature of the Persian court carpets of the time of Shāh ʿAbbās, was continued to the end of the 18th century. The cloud band became important on 16th-century carpets; it was employed with especial elegance and skill by Persian designers and perhaps most beautifully in Turkish court carpets, which owed much to Persian inspiration. The cloud band and knot motifs moved from Syrian textile design into Asia Minor with the Ottoman Turkish conquest in the 15th century and became typical of one group of 16th–17th century Turkish carpets.

Palmettes, a second major class of stylized motifs dominant in a considerable range of carpet designs from Asia Minor to India, originated in Assyrian design as stylizations of the palm, a symbol of vitalistic power that was often, if not always, associated with the Moon (Figure 79). Many of the almost uncountable variations that developed through the centuries continued to refer directly to the palm. As early as the 1st millennium BC, however, others



By courtesy of the Metropolitan Museum of Art, New York, gift of Mrs. Harry Payne Bingham, 1959; photograph, Otto E. Nelson—EB Inc.



Figure 78: Detail of a wool Persian arabesque carpet from Kerman, Iran, late 16th century. A system of double intersecting arabesque bands covers the field. In the Metropolitan Museum of Art, New York. Full size 6.05 × 2.49 m.

Figure 79: Detail of a wool Persian carpet from Kurdistan, Iran, late 18th century. Stylized palmettes dominate the field, which also includes motifs derived from the Chinese lotus blossom. In the Metropolitan Museum of Art, New York. Full size 6.96 × 2.69 m.

By courtesy of the Metropolitan Museum of Art, New York; gift of Joseph V. McMullan; photograph, Otto E. Nelson

derived from the lotus blossom, a complementary motif connected primarily with the fertility symbolism of the Sun. Still others involved the pomegranate, another fertility symbol, while yet another group presented the vitalistic emblem of the vine, this last design being built on the single leaf. The forms of these four main types of palmettes found in Oriental rug designs are directly descended from styles current in textile designs from the 4th century onward and are often modified by Chinese influences. The patterns in the 16th and early 17th centuries were beautifully and realistically elaborated, and blossoms such as the Chinese peony sometimes compete with the more stylized lotus. The lanceolate leaf, often associated with palmettes (especially in east Persian designs), is generally stylized. The chalice, fan, and half-palmette, all evolved from the palmette and used in Oriental rugs, were also used in 17th- and 18th-century European designs.

Naturalistic motifs

Outstanding among the more naturalistic plants are cypresses and blossoming fruit trees, symbolizing life eternal and resurrection, respectively. Willows and jasmine flowers are prominent in the Shāh ʿAbbās Vase carpet and tulips in Turkish court carpets. Many minor foliate and floral forms had no specific botanical identification, though they give a realistic effect (Figure 80). Naturalistic red or pink roses were widespread in European designs by the mid-16th century. Under European influence, they appeared in Oriental designs, particularly Persian, in the later 19th century.

The most important illustrative motifs, other than naturalistic plants, are those connected with the garden and the hunt: many small songbirds (in Persia, especially the nightingale); the pheasant (*feng-huang*), taken over from China and much favoured in the 16th century; occasionally the peacock; lions and a semi-conventional lion mask, sometimes used as the centre of a palmette; tigers; cheetahs; bears; foxes; deer of numerous species; goats, sometimes picturesquely prancing; the wild ass, a fleet prey; ferocious-looking Chinese dragons, and the gentle *ch'i-lin*, a fantastic equine also imported from China. Fish sometimes swim in pools or streams or are conventionally paired to suggest a shield, or escutcheon, in the borders



Figure 80: Indian wool floral carpet, Mughal, 17th century. The field is decorated with an allover pattern of naturalistic flowering plants. In the Metropolitan Museum of Art, New York. 4.27 × 2.01 m.

By courtesy of the Metropolitan Museum of Art, New York; photograph, Otto E. Nelson—EB Inc

of the carpet. Huntsmen, usually mounted, are the major human figures, though musicians are also depicted. Angels are occasionally present (see Figure 81).

The underlying theme of both the stylized and naturalistic vocabularies is nearly always fertility or abundance. The great Persian carpet of Ardabil (1539–40; Victoria and Albert Museum, London), for example, embodies a huge golden stellate medallion, developed from the multiple-pointed rosette that from time immemorial symbolized the Sun. At its centre are four lotus blossoms floating on a little gray-blue pool, which represents the source of rain in the heavens. The medallion thus symbolizes the two basic vitalizing elements—Sun and water. As proof of its magical potency, a complex system of tendrils and blossoms issues from it (see Figure 84).

Oriental versus European design

In Oriental carpet design, a flat surface pattern is always emphasized, even where small details are plentiful. European designs, however, tend more toward the illusionistic effects of painting, often using shading and picture-like compositions and incorporating architectural motifs and even portraiture. This tendency is particularly evident in French carpets of the 17th and 18th centuries.

SYMBOLISM OF OVERALL DESIGN

In addition to the symbolism inherent in individual motifs incorporated into the design of the carpet, the total design—indeed, the carpet itself—can be symbolic, as are some of the earliest Persian designs. The ultimate example is the Spring (or Winter) of Khosrow Carpet made for the audience hall of the Sāsānid palace at Ctesiphon (southeast of Baghdad) in the 6th century. The carpet has not survived, but, according to written records, it represented a formal garden with watercourses, paths, rectangular beds filled with flowers, and blossoming shrubs and fruit trees. Yellow gravel was represented by gold; and the blossoms, fruit, and birds were worked with pearls and different jewels. The outer border, representing a meadow, was solid with emeralds. Made of silk and measuring about 84 feet (25.6 metres) square, the carpet must have been overwhelmingly splendid when the great portal curtains of the hall were drawn back and the sun flooded the interior.

This dazzling carpet symbolized the divine role of the king, who regulated the seasons and guaranteed spring's return, renewing the earth's fertility and assuring prosperity. On another plane, it represented the Garden of Eden, a symbol of eternal paradise (the English word paradise is ultimately derived from the Persian word which means "walled park"). With its flowers, birds, and water, it symbolized deliverance from the harsh desert and the promise of eternal happiness.

This most sumptuous of fabrics made a profound impression on everyone, especially the Persians. For centuries it bewitched Persian imagination, becoming a legend in history, poetry, and art. Vain attempts at emulation were made by Oriental craftsmen for more than a millennium; and though its realistic depiction has disappeared, the

By courtesy of the Österreichisches Museum für Angewandte Kunst, Vienna; photographs, Eric Lessing—Magnum





Figure 81: *Naturalistic human figure and animal motifs.* Details of a Persian silk hunting carpet from Kashan, Iran, 16th century. In the Österreichisches Museum für Angewandte Kunst, Vienna. Full size 6.93 × 3.23 m. (Top) Winged figures from the border, possibly jinn or houris, seated amid flowering stems and birds in paradise. (Bottom) Scene from the field showing mounted huntsmen attacking a leopard.



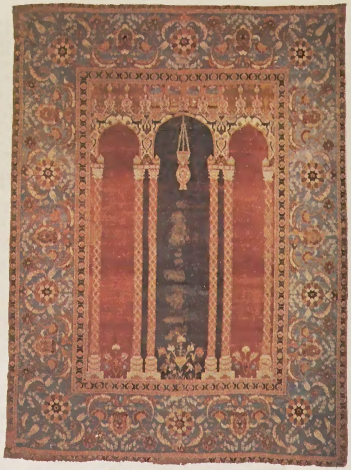


Figure 82: *Specialized rugs.*
(Left) Cruciform wool tabletop rug made in Cairo for export to Europe, Ottoman, 16th century. In the Museo d'Arte Sacra, San Gimignano, Italy. 2.60 × 2.30 m. (Right) Wool prayer rug (*namāzlik*) from Bursa, Turkey, Ottoman, 16th century. The field contains a *mihrāb*, or prayer niche, with a mosque lamp hanging in the central arch. In the Metropolitan Museum of Art, New York. 1.68 × 1.27 m.
By courtesy of (right) the Metropolitan Museum of Art, New York, gift of J.F. Ballard, photographs. (left) SCALA—Art Resource. (right) Otto E. Nelson—EB Inc.

Garden of Eden concept lingers on in Oriental designs. The garlands, vines, flowers, trees, animals, and beasts all strive to create a landscape, picturing hunting scenes or game, lakes with water birds, and often images of supernatural or celestial beings, such as jinn, houris, or a gathering of the blissful righteous at a banquet or dance (Figure 81). Accompanying verses support the image, lyrically extolling the carpet as a garden, for example, or a blooming meadow and comparing its beauty to that of the Garden of Eden.

Uses of rugs and carpets

Carpets originated in Central and Western Asia as coverings for beaten-earth floors. From time immemorial, carpets covered the floors of house and tent as well as mosque and palace. In the homes of wealthy Eastern families, floor coverings serve an aesthetic as well as a practical function. Rugs are often grouped in a traditional arrangement, partly to allow for simultaneous display; and the carpet's size and shape is determined by the intended place within that arrangement. There are usually four carpets. The largest, called *miān farsh*, usually measuring some 18 feet by 8 feet (5.5 by 2.5 metres), is placed in the centre. Flanking the *miān farsh* are two runners, or *kanārehs*, which are mainly used for walking and which measure some 18 feet by 3 feet (5.5 by 1 metre). The principal rug, or *kellegi*, averaging 12 feet by 6 feet (3.7 by 1.8 metres), is placed at one end of the arrangement of three carpets, so that its length stretches almost completely across their collective widths.

The prayer rug

The intended use sometimes determines both design and size, as in the prayer rug, or *namāzlik*. Because a Muslim must carry it everywhere, the prayer rug is relatively small (Figure 82, right). Design, naturally linked to religious imagery, is characterized by the *mihrāb*, or prayer niche (an imitation of the prayer niche in the wall of a mosque), the apex of which could be pointed toward Mecca. But other religious properties also appear, such as hanging lamps, water jugs, or "hand prints" to mark the place of the worshipper on the rug.

Until mid-17th century, Asian carpets imported into Europe were considered too precious to serve as permanent floor coverings. Placed on the floor only on church holidays or in an aristocrat's presence, they were otherwise used on the wall or to cover tables, benches, and chests; and, particularly in Italy, they were hung over balconies as decoration during festivals. Taking this European attitude into account, the Egyptian manufacturers created several unusual shapes and sizes for the European market: square, round, and cruciform carpets, obviously designed for tables rather than floors (Figure 82, left). During the 17th century, covering the entire floor with costly knotted carpets became fashionable. The mid-20th century witnessed a boom in antique-carpet prices that resulted in choicer pieces ending up back on the wall.

Table carpets

Oriental carpets frequently served many uses besides covering floors. They made handsome curtains, served as tribute money, and were frequently gifts of one state to another. They were used as blankets, canopies, coverings for tent openings, and tomb covers. They have also made excellent saddle covers and storage bags for use in tents. Such modest rugs were always close to the life of the people, who lavished care on them and into them wove life-protecting symbols. Other, more bizarre, uses have included assisting in the demise of Baghdad's last caliph—who in 1258 was wrapped in a carpet and beaten to death—and dramatically enhancing Cleopatra's introduction to Julius Caesar, when she stepped out of an unrolled rug. In less well-documented instances, they have assumed magical properties and taken flight.

Periods and centres of activity

Floor coverings of plaited rushes have been used at least since the 4th or 5th millennium BC, and rush weaving

in the Middle East had reached a high standard by medieval times.

ORIENTAL CARPETS

Oriental carpets are those made in Western and Central Asia, North Africa, and the Caucasus region of Europe. Rug design, in Western Asia at least, had gone beyond felt and plaited mats before the 1st millennium BC. A threshold rug represented in a stone carving (now in the Louvre) from the 8th-century-BC Assyrian palace of Khorsabad (in modern Iraq) has an allover field pattern of quatrefoils (four-leafed motifs), framed by a lotus border. Other Assyrian carvings of the period also show patterns that survive in modern designers' repertories.

Excavation of royal graves, dating from the 5th to the 3rd century BC, at Pazyryk, in the Altai Mountains of Southern Siberia, has uncovered the oldest known examples of knotting. The finds include various articles of felt with appliqué patterns and a superb carpet with a woollen pile, knotted with the Ghiordes, or Turkish, knot (Hermitage). The carpet, probably of Persian origin, measures $6 \times 6\frac{1}{2}$ feet (1.8×2.0 metres). The central field has a checkerboard design with a floral star pattern in each square. Of the two wide borders, the inner one shows a frieze of elk, the outer one a frieze of horsemen.

Knotting was not necessarily the only or even the most important method of carpet making. Felt carpets were used for a long time in Central and East Asia, as indicated by magnificent 1st-century-AD specimens from Noin Ula in Northern Mongolia (1st century BC to the 1st century AD; in the Hermitage) or those in the Shōsō-in (Japanese Imperial storehouse) in Nara near Osaka (before the 8th century). The costly rugs with figure motifs and gold mentioned by Greek and Arab writers may have been woven or embroidered and were probably exhibited on the wall as well as on the floor. The large carpet made in the 6th century for the Sāsānid palace in Ctesiphon (see above *Ornament and imagery: Symbolism of overall design*) is the most famous; but other Oriental courts, such as the caliphate at Baghdad (8th–13th century), also used valuable carpets.

Early Anatolian and Caucasian rugs

In the 13th, 14th, and 15th centuries, Asia Minor and the Caucasus produced coarse, vividly coloured rugs with stars, polygons, and often patterns of stylized Kūfic writing. A special group with simple, highly conventionalized animal forms was also woven; the most important of these carpets are represented by seven fragments of strong, repeating, geometric patterns in bold colours—red, yellow, and blue—found in the mosque of ʿAlāʾ ad-Dīn at Konya in Anatolia and now in the Museum of Turkish and Islāmic Art, Istanbul. They probably date from the 13th century. In the Staatliche Museen zu Berlin and in the Nationalmuseum at Stockholm are two primitive rugs, one, a highly conventionalized dragon-and-phoenix combat (Figure 83), the other, stylized birds in a tree. Both of these rugs are probably early 15th-century Anatolian.

Later, many rugs of finer weave, more delicate patterns, and richer colour—mostly geometric and possibly from Seljuq looms in Asia Minor—appeared in Europe. They were depicted by Flemish painters, such as Hans Memling, Jan van Eyck, and Petrus Christus, with such skill that the separate knots are sometimes visible. Many of these designs are repeated in the Bergama district of Asia Minor and the Southern Caucasus today, a complication when dating work.

Persia. Little is known about Persian carpet making before the 15th century, when the art was already approaching a peak. The Mongol invasion of the 13th century had depressed Persia's artistic life, only partly restored by the renaissance under the Mongol Il-Khan dynasty (1256–1353). Although the conquests of Timur Lenk (died 1405) were in most respects disastrous to Persia, he favoured artisans and spared them to work on his great palaces in Turkistan.

Under Timur's successor, Shāh Rokh (died 1447), art flourished, particularly carpets. Their production exclusively by palace workshops and court-subsidized looms gave them unity of style; and a sensitive clientele and lavish royal support guaranteed perfect materials and the highest skill: sheep were especially bred, dye plantations were cultivated like flower gardens, and designers and weavers could win court appointments. These conditions continued under the Ṣafavids (1501/02–1732).



Figure 83: Wool carpet with octagons containing a stylized dragon-and-phoenix combat motif, attributed to Anatolia, c. early 15th century. In the Staatliche Museen zu Berlin. 172 × 90 cm.
By courtesy of the Staatliche Museen zu Berlin, Islamisches Museum

In the 15th century the art of the book, which had long been considered the supreme artistic accomplishment and already had behind it centuries of superb achievement, reached a degree of elegance and sophistication unknown either before or since. The bindings, frontispieces, chapter headings, and, in the miniatures themselves, the canopies, panels, brocades, and carpets that furnished the spaces all received the richest and most elegant patterning. These beautiful designs were appropriated in various degrees by the other arts and account in no small measure for the special character of the court carpets of the period, the variety of colour, the ingenuity and imaginative range of pattern schemes, and the superlative draftsmanship that is both lucid and expressive.

Among the products inspired by book illumination were the Medallion carpets of northwest Persia, which consist of a large centre medallion connected with pendants or cartouches on the long axis and with quarter-section designs of the medallion in the corner areas. First used on ornamental pages and bindings of Persian books, on carpets this arrangement provided an effective centre and allowed several layers of designs to overlap because the medallions could cover multiple vine and flower patterns. The depiction of the latter motifs is more relaxed than their medieval rendering, and new motifs (inspired by painting) such as animals, humans, and landscapes began to be worked in.

A special court atelier, possibly located in Soltānīyeh, translated the most gorgeous illuminations into carpets. Among the 12 or so surviving examples are the world's most famous carpets, each a masterpiece of superb design, majestic size, purity and depth of colour, and perfection of detail. The best known are two carpets from the mosque at Ardabīl in eastern Azerbaijan, Iran, dated 1539–40. The better, skillfully restored, is now in the Victoria and Albert Museum (Figure 84); the other, reduced in size, is

The world's most famous Persian carpets

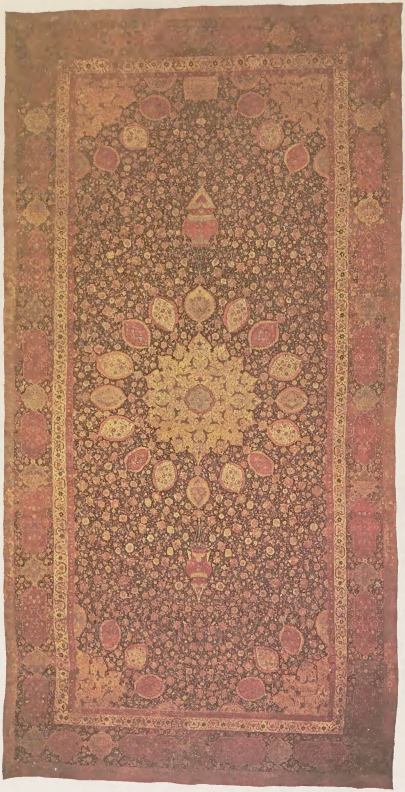


Figure 84: Wool and silk Persian medallion carpet from the mosque of Ardabīl (Iranian Azerbaijan), probably made in a workshop at Tabriz, Iran, dated 1539–40. A gold star medallion is centred on an indigo field of scrolling stems and blossoms. The medallion symbolizes the Sun; at its centre are four lotus blossoms in a pool, symbolizing the source of rain. In the Victoria and Albert Museum, London. 11.51 × 5.33 m.
By courtesy of the Victoria and Albert Museum, London

in the Los Angeles County Museum of Art. An extremely rich, intricate system of stems and blossoms covers a velvety, glowing indigo field, the whole dominated by a complex gold-star medallion. A near rival to the Ardabīl weaving is the Anhalt Carpet (possibly 19th century), named after a previous owner, the Duke of Anhalt, and now in the Metropolitan Museum of Art, New York City. An intricate star medallion dominates a brilliant yellow field covered with scrolling arabesques and fluttering cloud bands, framed by a scarlet border. One of the most beautiful of northwest Persian rugs is the "animal" carpet, half of which is in Kraków Cathedral, Poland, and half in the Musée des Arts Décoratifs. It has the same glowing scarlet and gold as the Anhalt Carpet but with more subtle halftones (buff on yellow, gray on taupe, brown on gray) and represents paradise more pictorially. Historically more important, and in beauty a rival of any, is the great "hunting" carpet in the Museo Poldi Pezzoli in Milan (Figure 85), inscribed: "It is by the efforts of Giyath-ud-Din Jami that this renowned carpet was brought to such perfection in the year 1521." A scarlet and gold medallion dominates a deep blue field, covered with an angular network of blossoming stems, across which hunters dash after their prey.

These carpets, in the opinion of many, represent the supreme achievement in the whole field of carpet designing. Nonetheless, other royal workshops were also producing many beautiful rugs. Particularly costly silk carpets with figure motifs (such as the silk hunting carpet in Vienna's Österreichisches Museum für Angewandte Kunst; see Figure 81) were woven in Kashan, Persia's silk centre. Smaller silk medallion carpets were also made there during the later 16th century, their designs mostly variations of the original medallion system. The court manufacture of Kashan also produced silk carpets with a decidedly royal style.

The distinctive rugs called Vase carpets (because of the flower vases in their designs) are generally thought to be Kerman. The pattern usually consists of several lattice systems with profuse blossoms and foliage. Many of these carpets survive as fragments; but only a scant 20 are intact, the finest of which is in the Victoria and Albert Museum. The rugs were apparently not for export but for court and mosque. Woven on a solid double warp, their boardlike stiffness holds them flat to the floor. In Iran they are still called "Shāh 'Abbās" carpets after the monarch of that name. The typically Persian style widely influenced carpets in Kurdistan and the Caucasus and also Indian court carpets, as well as embroideries from Bukhara.

Later in the 17th century, increasing luxury and wealth demanded the production of so many gold- and silver-threaded carpets that soon they were available in bazaars and exported to Europe, where more than 200 have been found. Some were made in Kashan, but many of the finest came from Isfahan. With their high-keyed fresh colours and opulence, they have affinities with European Renaissance and Baroque idioms. The Polish nobility ordered many gold-threaded rugs from Kashan, for Poland and Persia had close relations in the 17th century. Because there had been a rug- and silk-weaving industry using gold thread in 18th-century Poland, these imported Persian rugs, when first exhibited at the Paris exposition in 1878, were considered Polish, especially as nothing quite like them had at the time been found in Persia itself. They were accordingly dubbed "tapis Polonais," or Polish carpets, and the name has stuck. The type degenerated in the later 17th century, materials deteriorating, weaving coarsening, and designs muddling.

Herāt carpets

East Persian Herāt carpets, which were named after their centre of production and were characterized by their combination of a wine-red field and a border of clear emerald green with touches of golden yellow, became known in Europe as the typical Persian carpet. Many of the European artists of the period owned them, and Anthony Van Dyck and "Velvet" Brueghel (Jan the Elder), in particular, rendered them with complete fidelity in datable paintings. Indian princes also were enamoured of them and acquired them by plunder and purchase alike. Their popularity resulted in mass production with all its attendant deleterious effects, and the style finally expired in mediocrity.

Throughout 17th-century Persia, increasing refinement accompanied slackening inspiration. Silk carpets woven to surround the sarcophagus of Shāh 'Abbās II (died 1666) in the shrine at Qom (in Central Iran) were the last really fine achievements in Persian weaving. Even Orientalists have mistaken their finish for velvet; the drawing is beautiful, the colour varied, clear, and harmonious. The set is dated and signed by a master artist, Ni'mat Allāh of Jūsheqān.

At the end of the 17th century, nomads and town dwellers were still making carpets using dyes developed over centuries, each group maintaining an unadulterated tradition. Not made for an impatient Western market, these humbler rugs of the "low school" are frequently beautifully designed and are of good material and tech-



Figure 85: Detail of a Persian wool hunting carpet probably from Tabriz, Iran, dated 1521. Hunters and their prey are positioned symmetrically on a dark blue field covered with blossoming stems. The central motif is a scarlet and gold medallion; at its centre is an inscription cartouche signed Giyath-ud-Din Jami. In the Museo Poldi Pezzoli, Milan. 5.70 × 3.65 m.
SCALA, New York

nique. A great rug industry was developed in western Persia in the Soltānābād district; and from individual towns come beautifully woven rugs such as Sarūks, with their ancient medallion pattern; Serabands, with their repeating patterns on a ground of silvery rose; and Ferahans, with their so-called Herāti pattern—an allover, rather dense design with a light-green border on a mordant dye that leaves the pattern in relief. The earlier Ferahans (two are known, dated to the end of the 18th century) are on fields of dark lustrous blue with a delicately drawn open pattern. Later, Ferahans degenerated in colour, material, and design. "Low school" rugs maintained their standards down to the later 19th century, when insatiable Western demand ruined their artistry; but in the 20th century fine weaving in Persia has been somewhat revived.

Turkey. After the 16th century, Turkish rugs either followed Persian designs, indeed, were possibly worked by immigrant Persians and Egyptians, or followed native traditions. The former, made on court looms, displayed exquisite cloud bands and feathery, tapering white leaves on grounds of pale rose relieved by blue and emerald green. Turkish patterns embellished stately carpets, designed for mosques or noble residences, with rich, harmonious colours and broad, static patterns. They contrast with the lively, intricate Persian designs, in which primary, secondary, and tertiary patterns often interact with one another in subtle dissonances and resolutions.

Uşak carpets

Turkish styles are best illustrated by the carpets from Uşak in Western Anatolia, in which central star medallions in gold, yellow, and dark blue lie on a field of rich red. So-called Holbein rugs, similar to Caucasian carpets (see below), have polygons on a ground of deep red, dark green, or red and green; they often have green borders and conventionalized interlacing Kūfic script. Such a carpet is depicted in a portrait of Georg Gisze by the 16th-century German painter Hans Holbein the Younger—hence the name. Similarly, a handsome carpet pattern of interlacing yellow arabesques on a ground of deep red appears so often in the paintings of the 16th-century Venetian artist Lorenzo Lotto that they are called Lotto rugs. Carpets with a muted deep-red ground of wonderful intensity, patterned with small medallions, hail, perhaps, from Bergama. In the 17th century they developed into a type known as Transylvanian, so called because so many of them, particularly prayer rugs, were found in Transylvanian churches. They are nonetheless purely Turkish, with rich, quiet colour and sturdy designs. The majority are dominated by a fine red, though a few have faded to the colour of old parchment.

Turkish "bird carpets"

In the 17th century, the "bird carpet," or White Ushak, with conventionalized floral motifs suggesting birds, developed (Figure 86). Surviving examples are serenely beautiful, with fields of soft ivory and various discreet colours.

Eighteenth- and 19th-century "low school" rugs from Asia Minor continued the tradition of blending sober patterns and luxurious colour. Yürük "low school" rugs, made by nomadic Anatolian peoples like the Kurds, have attracted collectors because of their wide range of rich colours and the use of simple patterns, often geometric, that are organized in bold designs, frequently having a diagonal rather than a vertical emphasis. But the chief creations were prayer rugs, more plentiful among the Turks than among the other faithful. Handsome pieces were woven in Anatolia at Melas, Konya, Ladik, and Kirşehir, Ladik's being the most brilliant, both in colour and pattern. The most famous Anatolian prayer rugs came from Ghiordes and Kula, mostly in the 18th and 19th centuries; and in the United States they became the first passion of the collector. Regions such as Smyrna (İzmir) produced a great number of utility carpets for the West.

The Caucasus. Fine rugs were woven in the Caucasus

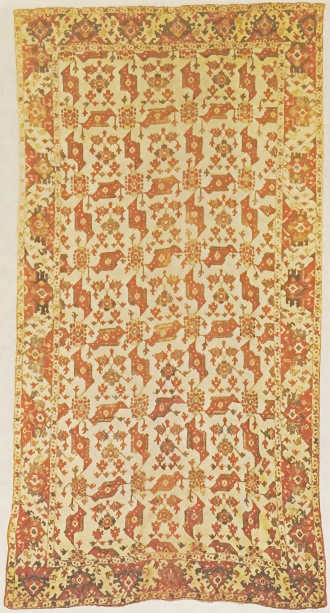


Figure 86: Wool "bird carpet," possibly from Uşak, Turkey, 17th century. The ivory white ground is patterned with an allover, stylized floral motif reminiscent of a bird. In the Metropolitan Museum of Art, New York. 4.44 × 2.31 m.

By courtesy of the Metropolitan Museum of Art, New York, gift of Joseph V. McMullan, 1963; photograph, Otto E. Nelson—EB Inc.

from the earliest times. During Persia's long political and cultural dominance, the magnificent carpets produced at the Persian court furnished models for the more ambitious Caucasian rugs, such as those woven for the local nobles, or khans. But the Caucasus has its own individual character; while it borrowed motifs from other areas, it completely transformed them by a furious vigour of design unequalled in the textile world. For example, although the Dragon carpets of Kuba continue medieval Persian motifs, the beasts, recognizable in the earliest Caucasian examples, are later stylized and enclosed in repeated rhomboid designs (Figure 87). This stylization process resulted perhaps from the combination of a taste for abstract design and the poverty of the region. The dense knotting required for curvilinear, natural designs is possible only with fine material, which the Caucasians could not afford. Their rugs, therefore, were of coarse weave, the Dragon carpets often having fewer than 80 knots per square inch (12 per square centimetre).

Caucasian "low-school" rugs

The "low-school" rugs are among the most individual and satisfactory. Their patterns are practically all geometric, densely juxtaposed, generally without organic connection or implied movement; but they are clear, ingenious, and entirely suitable for floor decoration. More recent examples seem a little dry in colour; but rugs woven by the Kazakhs, Sarüqs, and other nomads, are sometimes of flaming brilliance; and the older rugs from Dagestan, Kuba (both west of the Caspian Sea), and Shirvan (on the borders of Iraq and Iran) are done in beautifully clear, discreet, and well-balanced tones.

Shirvan *kilims*, or tapestry rugs, with their broad horizontal stripes, have bold motifs assembled in harmonious colours.

Another type of flat-stitch carpet, brocaded with a mass of loose threads at the back, comes from the Shemakha region (in Azerbaijan). It has mistakenly been called cashmere because of its superficial resemblance to cashmere shawls. The design usually embodies large, beautifully articulated mosaic-tile patterns in rich and sober colours.

Turkistan. The carpets of western Turkistan (wrongly called Bokhara carpets) are made by nomadic Turkmen tribes. Few extant examples are more than 100 years old, though similar rugs have almost certainly been made for centuries. Many older pieces are not intended as floor coverings. Some, called *jovāls*, are bags (about 5 by 3 feet; 1.5 by 1 metre) for storage in tents. Some are saddlebags consisting of two squares of about 2 feet (0.6 metre), joined together. There are also long tent bands (called *kibitkas*) about 1 foot (0.3 metre) wide and perhaps 60 yards (55 metres) long, used for decorating large tents, and rugs used as hangings for tent doorways. Small, squarish rugs and larger ones of about 10 by 7 feet (3 by 2 metres) seem later and were made perhaps mainly for sale. Turkmen carpets have woollen warp, weft, and pile, two lines of weft, and either Sehna or Ghiordes knots. Except for the Baluchi, Turkmen rugs are characterized by a dark-red colouring and geometric designs. After the predominant red, the chief colours are blue, white, and a natural black wool toning pleasantly to brown. The characteristic design is the octagon, or elephant's foot, arranged in rows and columns, often with diamond-shaped figures in between. Door hangings have cross-shaped panelling, smaller pieces a rectangular diaper (allover pattern). Woven end webs and tassels are used freely as embellishments. Carpets made by Turkmen tribes are the Tekke, Yomut, Afghan, Sarük, Ersar, Beshir, and Baluchi.

Characteristics of Turkmen rugs

Chinese Turkistan. The oldest surviving Chinese Turkistan rugs date perhaps from the 17th century. Most have a silk pile and some metal and gilded thread, with floral, Persian-influenced patterns showing distinct Chinese treatment. Later carpets are loosely woven with the Sehna knot and have a wool or, more rarely, silk pile and a cotton warp. Eighteenth-century examples have rich, dark

The Textile Museum Collection, Washington, D.C.; photograph, Otto E. Nelson—EB Inc.



Figure 87: Detail of a wool Kuba Dragon carpet, probably from Karabagh or Shirvan in southern Caucasia, 17th century. The dragon, enclosed in an ogee lattice intersected by palmettes and blossoms, is derived from Chinese motifs through medieval Persian models. In the Textile Museum, Washington, D.C. Full size 5.34 × 2.39 m.



Figure 88: Chinese Turkistan three-medallion wool carpet from Khotan (Hot'ien), Sinkiang Uigur Autonomous Region, China, 19th century. The top and centre medallions contain a pomegranate branch and vase motif. The corner filling is a Turkish form of the Chinese cloud scroll. The guard stripe has a stylized Chinese wave pattern, the centre band of the main border a swastika meander. In the collection of Vojtech Blau, New York. 3.71 × 1.83 m.

In possession of Vojtech Blau, New York City; photograph Otto E. Nelson—EB Inc.

colouring, which became brighter in the 19th century and, at last, excessively crude. There are two main types of design. The Medallion carpet usually has three squarish medallions placed down the centre and, almost invariably, one border with a conventional Chinese pattern of foam-crested waves (Figure 88). This design is generally called Samarkand in the trade, though the rugs themselves come from Kashgar, Khotan, Yarkand (modern Su-fu, Hot'ien, So-ch'e in China's Sinkiang region). The Five-blossom carpet has a floral diaper with groups of five blossoms. The colouring is often red and orange with a little clear blue.

Egypt. Egyptian carpets used to be called Damascus carpets but are now termed Mamlūk, after the Muslim dynasty (1250–1517) that subsidized their manufacture, or Cairene rugs after Cairo, the city in which they were made.

Knotting was thought to have reached Egypt from Asia Minor, but early Cairene carpets differ from Anatolian ones: they are knotted with the Persian knot, and their colours are red, yellow green, and light blue, applied evenly in inner field and border. Moreover, Egyptian designs concentrate on the centre, subordinating surrounding motifs. Border designs match rectangular with square panels. Although filled with plants, the carpet designs seem geometric (Figure 89). The heyday of these rugs occurred during later Mamlūk rule, when there was extensive export.

India. Carpets are less important in India than elsewhere in Asia because the climate makes knotted floor coverings unsuitable. As an art, carpet weaving was brought from Persia by the 16th- and 17th-century Mughal emperors, who subsidized the manufacture of beautiful rugs with an almost silken sheen. Although Indian artistry was influenced by Southern and Eastern Persian carpets, it maintained a native taste for pictorial representation (Figure 90).

Mughal carpets

The carpets made for the courts of the grand Mughals were of extravagant and luxurious beauty. Expense was ignored, and a series of carpets was made with 600 to 1,200 knots per square inch (95 to 190 per square centimetre). Special carpets were of even finer weave, 2,100 knots per square inch (325 per square centimetre; Metropolitan Museum of Art). For the palace of Shāh Jahān (died 1666) a set of rugs was made from the most precious wool, imported from Kashmir and remote Himalayan valleys. But because the sources of the art were in imitation, not in the life roots of the people, these wonderful fabrics never reached the artistic height that characterized many periods of Persian weaving. Once established, the rug industry continued, becoming a jail industry, particularly in the Punjab. Designs degenerated, and good wool was difficult to obtain. Later Indian carpets are thus mostly inferior to Persian work.

China. The rugs of China are recognizable by their characteristic Chinese ornament. Of coarse texture, they are Sehna-knotted on a cotton warp. The pile is thick, smooth-surfaced, and "sculptured" so as to form a furrow at the pattern's contours. Yellow predominates, sometimes intentionally, sometimes as the result of the fading of red and orange. Blue and white are commonly used; but true red, brown, and green are rarely seen.

Some carpets have repeating plant scrolls; others scattered flowers and various Chinese symbols. Frets, or key designs, often decorate the border.

Chinese Pillar carpets

Peculiar to China are Pillar carpets, designed so that when wrapped around a pillar the edges will fit together to form a continuous pattern, usually a coiling dragon (Figure 91). Small mats and seat covers are also common. Chinese rugs are virtually impossible to date, since they vary little with time. Many large carpets have been made for export in the 20th century.

By courtesy of the Metropolitan Museum of Art, New York, gift of George Blumenthal; photograph, Otto E. Nelson—EB Inc.



Figure 89: Cairene wool carpet from Egypt, 16th century, Mamlūk period. The field features a star medallion centred in a geometrically designed ground, covered with stylized forms of the papyrus and other plants. In the Metropolitan Museum of Art, New York. 2.41 × 2.17 m.



Figure 90: Indian wool pictorial carpet, Mughal, late 16th or early 17th century. The field design resembles Mughal painted miniatures, or illuminated manuscripts. In the Museum of Fine Arts, Boston. 2.4 × 1.5 m.

By courtesy of the Museum of Fine Arts, Boston, gift of Mrs. Frederick L. Ames in the name of Frederick L. Ames

WESTERN CARPETS

Spain. Spain's close ties with Islām after the 8th century made it quick to accept and produce knotted carpets. Early examples of the unusual Spanish knot suggest manufacture as early as the 12th century, but not until the 15th century do enough examples remain to allow grouping of work. Many designs imitate Anatolian forms; others, with coats of arms or Christian emblems, indicate purely European origin. During the 16th century, Renaissance influence was prevalent. The manufacturing centres were Cuenca, Alcaraz, and possibly Almería. The knotted carpet lost ground during the 18th century; and native work, known as Alpujarra (after the district), is embroidered or done in uncut weft-loop technique.

France. In France, too, the stimulus for the production of knotted carpets came from the East; but the designs of the rugs were inspired by contemporary French decoration rather than Oriental carpet design. Jean Fortier and Pierre Dupont won fame knotting pieces in the Hospice de la Savonnerie at Chaillot, which was converted from a soap factory to a carpet factory in the early 17th century. "Savonnerie" became a mark of distinction in French carpets, reaching a zenith during the later 17th century with Louis XIV's immense order for Versailles. Combined since 1826 with the Gobelins factory, the firm still operates. Thick and strong, these carpets consist of a woollen pile on a mostly linen warp. During the 18th century and afterward, many tapestry-woven carpets were made at Aubusson as well as at other tapestry factories. Even though their production has not been confined to that city, they are known as Aubusson carpets.

Savonnerie and Aubusson carpets

The European concept of carpet design, as distinguished from the Oriental concept, is most explicit in the Savonnerie carpets, in which three-dimensional compositions complement architecture, and even portraits are reproduced (Figure 92). The style of such carpets is best seen in sketches of rug design made by Charles Le Brun for Louis XIV (mostly in the Mobilier National in Paris).

United Kingdom and Ireland. The growth of a native craft in the United Kingdom soon followed on the introduction of carpets from Turkey, though 16th- and 17th-century intact specimens number only about a dozen. They are characterized by a hemp warp and weft, medium-fine woollen pile, and the Ghiordes knot. The background usually is green, and there are so many shades of the other colours that the entire number of tints is greater than in Oriental carpets. The designs can be divided into two groups. In the first are typically English patterns resembling contemporary embroidery, often with heraldic devices and dates. The oldest specimen, dated 1570, belongs to the Earl of Verulam. In the second group are many pieces of carpet knotting—called at the time "Turkey work"—imitating Oriental designs and made to cover chairs and stools. As the demand for carpets increased in the 18th century, factories were established at Paddington, Fulham, and Moorfields, near London, and at Exeter and Axminster in Devon. Axminster worked on well into the 19th century, when it merged with the Wilton Royal Carpet Factory Ltd. at Wilton, Wiltshire, which still operates. The industry dwindled and almost disappeared with the advent of mechanization until about 1880. The craft was revived by the English artist and poet William Morris. Later in the 19th century, a factory opened in Donegal, Ireland; and during the 20th century, many small rugs have been knotted by handicraft societies.

Axminster carpets

Scandinavia. Scandinavian work is similar in concept despite national differences of colour and motif. Abundant

The Textile Museum Collection, Washington, D.C.; photograph, Otto E. Nelson—EB Inc.



Figure 91: Chinese wool Pillar carpet, late 19th century. When the rug is placed around a pillar, the dragon becomes continuous, and the animal masks at the top form a capital. Chinese cloud motifs and Buddhist symbols cover the field. At the base, a mountain rises from ocean waves. In the Textile Museum, Washington, D.C. 2.41 × 1.23 m.

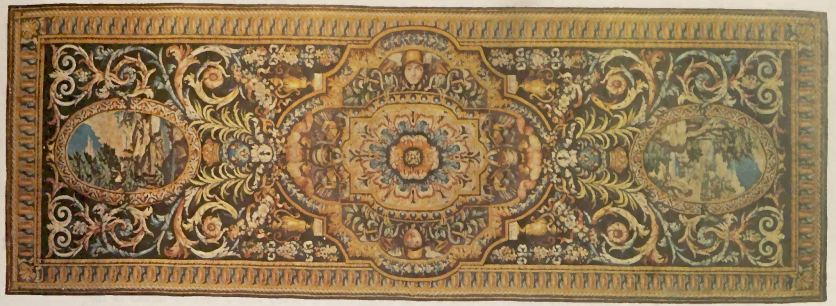


Figure 92: Savonnerie wool carpet made at Chaillot, France, 17th century. A predominately floral arabesque design covers the field, at either end of which are medallions with landscape scenes. The centre motif is a Baroque cartouche. In the collection of Mr. and Mrs. Charles B. Wrightsman, 3.0 × 9.1 m.
Collection of Mr. and Mrs. Charles B. Wrightsman, photograph: the Metropolitan Museum of Art, New York

handmade products include floor coverings, coverlets, and upholstery for benches, chairs, stools, and pillows. Techniques dating from the Vikings (and probably imported by them from Turkey) are continued in Swedish and Finnish rugs, called Rya rugs. Knotted work includes pieces with pile on either side, many Ghiordes on three warps, and braided and woven patchwork carpets with interwoven strips. Geometric designs, rooted in the native arts, are common, appearing, for example, in opulent "wedding carpets." Design was also influenced by Dutch tapestry flower motifs.

Eastern Europe. Knotted Mazovian rugs of East Prussia show the strongest Oriental influence, though at the same time they are deeply rooted in peasant traditions. Many other textiles untouched by west European influence, however, came from southeast Poland, the Ukraine, and southern Russia; some are characterized by ancient textile motifs (such as simple stripes) and forceful colour harmonies, others by geometric designs resembling those of the Orient. *Kilims*, or tapestry-woven carpets, are common in those areas, as they are in the Balkans. In Romania, government promotion and the interest taken by contemporary artists in folk idiom have stimulated modern production during the 20th century.

European folk carpets. Carpet making is so widespread in European folk art that it probably would have developed even without stimulus from the Orient. The most varied techniques are represented in these tradition-bound products the designs of which remained unchanged for generations. The work includes floor coverings, chest covers and bedcovers, and draperies, most of modest size (or pieced together) and many made in sets. The colour scheme is very limited, for even the raw materials were homemade. Machine-made carpets in the later 19th century quickly engulfed home products, but a conscious revival and renewal followed in the 20th century.

North America. The technique of knotting has not been used by the Indians, but many tribes have been making flat-woven floor rugs and blankets since the earliest days of their known history. Before sheep were introduced in the 16th century and wool became dominant, the principal material was cotton, together with various fibres and dog's hair. Indian designs are traditionally abstract, making much use of stripes and a zigzag, or "lightning," motif. The colours are black, white, yellow, blue, tan, and red, the latter often dominant. Among the most skillful carpet makers are the Pueblo and Navaho tribes.

North American Indian rugs

Rugs were made by the colonists in a variety of techniques: knitting; crocheting; braiding thin strips of material into small squares and then sewing them together; and embroidering on a coarse-woven foundation. Hooking (drawing material through a canvas foundation) began around the turn of the 18th century and became very popular; early examples have crude floral, geometric, or animal designs and are very colourful. No knotted carpets were manufactured by the early settlers. In 1884, however, a factory established in Milwaukee (and later moved to New York City) began to weave carpets in traditional European designs. During the 1890s a branch of the English Wilton Royal Carpet Factory made Axminsters at Elizabethport, New Jersey; and a few beautiful, flat-woven carpets in French Baroque and Neoclassical designs were produced around the turn of the century by a tapestry factory in Williams Bridge, New York. After this, machine weaving, which began in the United States in the late 1700s, gradually displaced hand weaving. (Ed.)

TAPESTRY

Historically almost any heavy material, handwoven, machine woven, or even embroidered, used to cover furniture, walls, or floors or for the decoration of clothing, has been called tapestry in popular usage. Since the 18th and 19th centuries, however, the technical definition of tapestry has been narrowed to include only heavy, reversible, patterned or figured handwoven textiles, usually in the form of hangings or upholstery fabric. Tapestry traditionally has been a luxury art afforded only by the wealthy, and even in the 20th century large-scale handwoven tapestries are too expensive for those with moderate incomes.

Tapestry sets

Tapestries are usually designed as single panels or sets. A tapestry set is a group of individual panels related by subject, style, and workmanship and intended to be hung together. The number of pieces in a set varies according to the dimensions of the walls to be covered. The designing of sets was especially common in Europe from the Middle Ages to the 19th century. A 17th-century set, the "Life of Louis XIV," designed by the king's painter Charles Le Brun, included 14 tapestries and two supplementary panels. The number of pieces in 20th-century sets is considerably smaller. "Polynesia," designed by the modern French painter Henri Matisse, for example, has only two pieces, and "Mont-Saint-Michel," woven from a cartoon by the contemporary engraver and sculptor Henri-Georges Adam, is a triptych (three panels). Until the 19th century, tapestries were often ordered in Europe by the "room" rather than by the single panel. A "room" order included

not only wall hangings but also tapestry weavings to upholster furniture, cover cushions, and make bed canopies and other items. Most Western tapestry, however, has been used as a type of movable monumental decoration for large architectural surfaces, though in the 18th century, tapestries were frequently encased in the woodwork.

In the West, tapestry traditionally has been a collective art combining the talents of the painter, or designer, with those of the weaver. The earliest European tapestries, those woven in the Middle Ages, were made by weavers who exercised much of their own ingenuity in following the cartoon, or artist's sketch for the design.

Though he followed the painter's directions and pattern fairly closely, the weaver did not hesitate to make departures from them and assert his own skills and artistic personality. In the Renaissance, tapestries increasingly became woven reproductions of paintings, and the weaver was no longer regarded as the painter's collaborator but became his imitator. In medieval France and Belgium, as well as now, a painter's work was always executed in tapestry through the intermediary of the weaver. Tapestry woven directly by the painter who created it remains an exception, almost exclusive to ladies' handiwork. This section covers the materials, techniques, and history of tapestry making.

Materials

Wool has been the material most widely used for making the warp, or the parallel series of threads that run lengthwise in the fabric of the tapestry. The width-running, weft, or filling threads, which are passed at right angles above and below the warp threads, thereby completely covering them, are also most commonly of wool. The advantages of wool in the weaving of tapestries have been its availability, workability, durability, and the fact that it can be easily dyed to obtain a wide range of colours. Wool has often been used in combination with linen, silk, or cotton threads for the weft. These materials make possible greater variety and contrast of colour and texture and are better suited than wool to detail weaving or to creating delicate effects. In European tapestry, light-coloured silks were used to create pictorial effects of tonal gradation and spatial recession. The sheen of silk thread was often used for highlights or to give a luminous effect when contrasted to the dull and darkly coloured heavier woollen threads. In 18th-century European tapestries, silk was increasingly used, especially at the Beauvais factory in France, to achieve subtle tonal effects. Most of the Chinese and Japanese tapestries have both warp and weft threads of silk. Pure silk tapestries were also made in the Middle Ages by the Byzantines and in parts of the Middle East. Wholly linen tapestries were made in ancient Egypt, while Copts, or Egyptian Christians, and medieval Europeans sometimes used linen for the warp. Cotton and wool were employed for pre-Columbian Peruvian tapestries as well as for some of the tapestries made in the Islāmic world during the Middle Ages. Since the 14th century, European weavers have used gold and silver weft threads along with wool and silk to obtain a sumptuous effect. These threads were made of plain or gilded silver threads wound in a spiral on a silk thread.

Techniques

Tapestry is first of all a technique. It differs from other forms of patterned weaving in that no weft threads are carried the full width of the fabric web, except by an occasional accident of design. Each unit of the pattern or the background is woven with a weft, or thread of the required colour, that is inserted back and forth only over the section where that colour appears in the design or cartoon. As in the weaving of plain cloth, the weft threads pass over and under the warp threads alternately and on the return go under where before it was over and vice versa. Each passage is called a pick, and when completed the wefts are pushed tightly together by various devices (awl, reed, batten, comb, or serrated fingernails in Japan). The weft threads so outnumber the warps that they conceal them completely. The warps in a finished tapestry appear only as more or less marked parallel ridges in the texture, or grain of the fabric, according to their coarseness or fineness.

Weaving tapestry

The thickness of the warp influences the thickness of the tapestry fabric. In Europe during the Middle Ages, the thickness of the wool tapestry fabric in such works as the 14th-century "Angers Apocalypse" tapestry was about 10 to 12 threads to the inch (five to the centimetre). By the 16th century the tapestry grain had gradually become finer as tapestry more closely imitated painting. Known for the regularity and distinctness of its tapestries, the royal French tapestry factory in Paris known as the Gobelins used 15 to 18 threads per inch (six to seven per centimetre) in the 17th century and 18 to 20 (seven to eight) in the 18th century. Another royal factory of the French monarchy at Beauvais had as many as 25 or even 40 threads

By courtesy of the Mobilier National, Paris; photograph, Visages de France

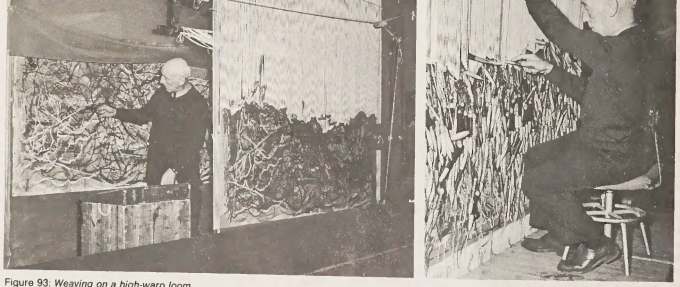


Figure 93: *Weaving on a high-warp loom.*
(Left) Weaver indicating an area on the cartoon that corresponds to the portion of the tapestry (right foreground) he has woven. (Right) Weaver, working on the reverse side of the tapestry, pulls a set of warps forward and passes the bobbin behind it. (Photographs taken at the Gobelins factory, France.)

per inch (10 to 16 per centimetre) in the 19th century. These excessively fine grains make the fabric very flat and regular, tending to imitate the canvas of a painting. The grain of 20th-century tapestry approximates that used in 14th- and 15th-century tapestry. The Gobelins factory, for instance, uses 12 or 15 threads per inch (five or six per centimetre).

In many 20th-century tapestries a finer grain is contrasted with the effects of a heavier weave. The grain of silk tapestries, of course, is much finer than those made of wool. It is not uncommon for the silk tapestries of China to have as many as 60 warp threads per inch (about 24 per centimetre).

Where the weft margin of a colour area is straight and parallel to the warps, it forms a kind of slit, or *relais*, which may be treated in any of five different ways. First, it may simply be left open, as in Chinese silk tapestries, which are called *k'o-ssu* (cut silk) for that reason. Second, it may be left open on the loom but sewed up afterward, as in European tapestries from the 14th to the 17th centuries and also in some later types. Third, the weaver may dovetail his wefts, passing from one side and from the other in turn over a common warp. This may be either "comb" dovetailing—single wefts alternating—or "sawtooth" dovetailing—clusters first from one side, next from the other. Dovetailing has the double disadvantage of making the fabric heavier at that point and of blurring the outline. Persian weavers of the 16th century developed a successful variant in silk tapestry rugs whereby a black outline weft was dovetailed over two warps—one of each of the adjacent colour areas—effectively hiding the coloured wefts in the compacting of the weave and providing a strong clear image. The same device is found in pre-Columbian Peru.

The fourth treatment—interlocking—was introduced in the Gobelins factory in the 18th century. Here wefts of juxtaposed colour segments are looped through each other between the two warps that mark, respectively, the margin of each colour. This technique produces a continuous surface of even weight that was prized by the French weavers because the resultant effect more closely approximated that of painting.

A curious variant of these weaving techniques is achieved when between every two rows of wefts there is a weft that runs the full width of the tapestry, thereby making the fabric solid. This technique, if strictly classified, would be called brocade weaving, but the principle is that of tapestry, with the cloth insert subordinate. Rarely used, the technique was employed in Japan in the 7th and 8th centuries, in eastern Persia in the 10th century, and in pre-Columbian Peru.

Twill technique of weaving

Instead of the plain-cloth method of weaving usually used in making tapestries, a twill technique can be used. In this type of weave the weft is floated over two or more warps, then under one or more warps, with this underpassage shifting always one to the right or left, thereby making a diagonal ribbing. As far as can be determined, this technique first appeared in medieval Persia and from the 17th century on was especially used in the Iranian provinces of Khorāsān and Kermān to make shawls of goat's hair or wool. It is also used to make the famed Kashmir shawls and, along with many other crafts, was probably introduced into Kashmir from Persia, in the 16th century. In contemporary European tapestries this technique, usually called eccentric weaving, occasionally has been used in making some of the experimental abstract hangings of the later 20th century.

European tapestry may be woven on either a vertical loom (high-warp, or *haute-lisse* in French) or a horizontal loom (low-warp, or *basse-lisse*). In early high-warp looms the warps were attached to a beam at the top, and groups of warp threads were weighted at the bottom. The weft was beaten up (*i.e.*, pushed) toward the top as the weaving progressed. High-warp looms of this type are pictured on ancient Greek vases. In later high-warp looms the vertical frame has heavy uprights holding a horizontal roller at top and bottom, on which the warps are stretched. Each warp passes through a loop of cord (the *lisses*), and the loops encircling the warps that correspond to uneven numbers are fastened to one slender cylinder; those to the even-numbered warps are fastened to another cylinder. Both cylinders are above the weaver but within reach so that he can pull forward first with one, then with the other set of warps (*i.e.*, form the shed) in order to pass his bobbin behind them. The bobbin (broche) is a short, pointed, slim cylinder of polished wood on which the weft yarn is wound (Figure 93).

The low-warp loom, on the other hand, has the rollers on the same level at table height so that the warps stretched between them are horizontal. To leave the weaver's hands free, the warps are attached to two slats, or poles, each of which is connected with a treadle so that the weaver's foot depresses the odd-numbered or even-numbered series of warps to form a passageway for the bobbin, called a shuttle on the low-warp loom. The cylinders in both instances serve to roll up the finished portion and unroll a further length of unwoven warps so that the section in process is always taut and in a convenient relation to the weaver (Figure 94). At both types of loom the weaver works from the back side, that is, he weaves the tapestry on the wrong side. He has, however, a hand mirror, which he puts through the unwoven warps holding it to reflect the right

By courtesy of the Mobilier National, Paris; photograph, Visages de France



Figure 94: *Weaving on a low-warp loom.*
(Left) Weaver depresses the treadles with his feet, thereby controlling the space between warps through which the shuttle, held in his right hand, will be moved. (Right) Working on the reverse side of the tapestry with the cartoon positioned under the warps, the weaver passes the shuttle between the warps. (Photographs taken at the Beauvais factory, France.)

side of the portion in process. While the high-warp weaver can examine his finished work directly by walking around to the other side of his loom, the low-warp worker has to tilt up his frame.

Of the two techniques, low-warp is more commonly used. Of the great European tapestry works only one, Gobelins, has traditionally used high-warp looms. Several weavers can work simultaneously on either kind of loom. Depending on the complexity of the design and the grain or thickness of the tapestry texture, a 20th-century weaver at the Gobelins can produce 32 to 75 square feet (three to seven square metres) a year.

The medieval cartoon

In Western tapestry the medieval cartoon, or preparatory drawing, was usually traced and coloured by a painter on a canvas the size of the tapestry to be woven. At the end of the 15th century the weaver probably wove directly from a model, such as a painting, and consequently copied not a diagrammatic pattern but the original finished work of the painter. At the beginning of the 17th century there arose a clear distinction between the model and the cartoon. The model was the original reference on which the cartoon was based. Cartoons were rapidly and freely used and were often copied.

More than one tapestry can be woven from a cartoon. At the Gobelins factory, for instance, the 17th-century "Indies" tapestry set was woven eight times, remade, and slightly altered by the late Baroque painter François Desportes (1661–1743); these cartoons were woven several more times during the 18th century.

The border of a cartoon tended to be redesigned every time it was commissioned, since each patron would have a different heraldic device or personal preference for ornamental motifs. Borders were frequently designed by an artist different from the one who conceived the cartoon for the central narrative or principal image. As an element of tapestry design, however, borders or frames were important in European tapestry only from the 16th to the 19th century. Medieval and 20th-century tapestries seldom use this device, which emphasizes the idea of the tapestry as a reproduction of or substitution for a painting.

Contemporary cartoons

A fully painted cartoon requires much of the painter's time and is tedious to make. In the 20th century, therefore, other solutions have been adopted. The cartoon may be a photographic enlargement of a fully painted model or, more simply, a numbered diagrammatic drawing. The latter type of cartoon was worked out by the famous French tapestry designer Jean Lurçat during World War II. In this method each number corresponds to a precise colour and each cartoonist has his own range of colours. The colours are not indicated in a photographic enlargement, but the weaver refers to a small colour model provided by the painter and from it makes a selection of wool samples.

The high-warp weaver has the full-size cartoon, which he follows as it hangs beside or behind him. The low-warp worker has the cartoon laid under the warps, so he follows it from immediately above. In both cases the main outlines are drawn with ink on the warps after they have been mounted, or attached to the loom. The design is executed, in all European work since the Middle Ages, at right angles to the loom, so that in the finished hanging the warps usually run horizontally rather than vertically as they ran on the loom. Though in certain pieces the warps run vertically, it is aesthetically advantageous for the tapestries to be executed horizontally, since the warp ribbing tends to create a texture more or less reinforced by linear shadows, which, if vertical, sever the design but if horizontal bind it into continuity. Practically, however, horizontal warps are disadvantageous, since the horizontal slits made in weaving will pull apart more rapidly than vertical slits because of the weight of the hanging.

Periods and centres of activity

ANCIENT WESTERN WORLD

Examples of tapestry weaving from the ancient world are so isolated and fragmentary as to make it uncertain either when or where the art originated. The earliest known tapestry weaving was done in linen by the ancient Egyptians between 1483 and 1411 BC. Preserved by the dry desert climate of Egypt, three tapestry fragments were found in the tomb of Thutmose IV. Two of the fragments have cartouches of Egyptian pharaohs, and the third is a series of hieroglyphs. In the tomb of Tutankhamen (*c.* 1323 BC), a robe and glove woven by the tapestry technique have also been found.

Although no examples remain, writers of antiquity are unanimous in proclaiming the magnificence of Babylonian and Assyrian tapestries. Some scholars have speculated that the ancient Egyptians learned the art of tapestry from the ancient peoples of Mesopotamia. During that period when the few preserved Egyptian tapestry fragments were made, Mesopotamian ideas, techniques, and, perhaps, craftsmen were entering Egypt. These scholars conjecture that, since tapestry weaving did not occur in quantity again in Egypt until the 4th century AD, it is likely that the craft was not indigenous.

Tapestry weaving continued to flourish in western Asia in the 1st millennium before Christ. Fragments of wool tapestries dating from the 4th or 3rd century BC have been found in graves in Ukraine near Kerch in the Crimean peninsula. The ornamental motifs of these fragments are of a widely diffused Hellenistic style that was especially prevalent in Syrian art at the time. Another fragment showing close Syrian connections is a piece of silk tapestry dating about 200 to 500 years later and found in China at Lou-lan in Sinkiang Uighur Autonomous Region. Other fragments have been found in Syria at the archaeological sites of Palmyra and Doura-Europus. If climatic conditions for textile preservation in the Middle East had been more favourable, it might be possible to theorize that Syria was a great centre of tapestry weaving, especially at the start of the Christian Era.

Tapestry in ancient Greece and Rome

There are literary descriptions of the making of tapestry in ancient Greece and Rome. In the *Odyssey*, Homer (8th century BC?) describes Penelope working on a tapestry that was unravelled each night as she waited for Odysseus. The Roman poet Ovid (43 BC–AD 17) in the *Metamorphoses* describes the tapestry looms used by Minerva and Arachne in their mythological weaving contest. During the period of the empire the Romans apparently imported a considerable number of the tapestries used in their public buildings as well as in the homes of the wealthy. Since the Latin terms referring to tapestry and weaving are Greek in origin, it is generally supposed that the art of tapestry making was taught to the Romans by the Greeks.

EASTERN ASIA

Called *k'o-ssu* (cut silk), tapestry has long been produced in China, traditionally being made entirely of silk; Chinese tapestries are extremely fine in texture and light in weight. The weave is finished perfectly on both sides so

By courtesy of (left) the Musee Historique des Tissus, Lyon, France, (right) the Metropolitan Museum of Art, New York, gift of Ellis G. Seymour, 1925

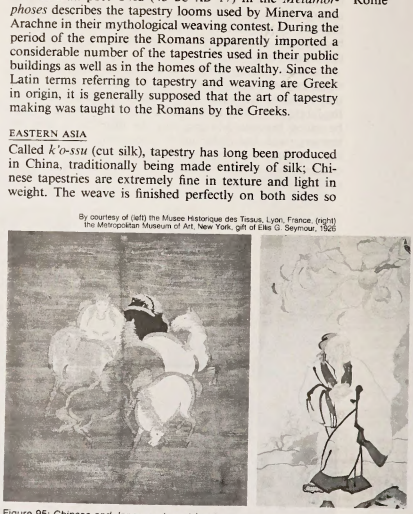


Figure 95: *Chinese and Japanese tapestries.*
(Left) "Horses," Japanese *tsuzure*, Edo period, late 18th century. In the Musée Historique des Tissus, Lyon, France. 70 × 65 cm. (Right) "T'ung Fung Stealing the Peaches of Longevity," Chinese *k'o-ssu*, Ming Dynasty (1368–1644). In the Metropolitan Museum of Art, New York. 1.17 × 0.61 m.

that the tapestries are reversible. The warps are vertical in relation to the pattern, rather than horizontal as in European weaving. Sometimes the weaver uses metal threads to make his hangings more sumptuous or highlights the design by painting, although this is not considered a commendable expedient.

Many *k'o-ssu*, such as "T'ung Fung Stealing the Peaches of Longevity" (Figure 95), imitated paintings and were mounted on scrolls or album leaves in the same manner as the pictures they copied. Tapestries to cover large wall surfaces, such as the *k'o-ssu* (seven feet three inches by five feet nine inches; 2.2 by 1.75 metres) of "Feng-huang in a Rock Garden" (late Ming period, Metropolitan Museum of Art, New York), were usually brighter in colour, heavier in texture, and frequently woven with metal threads. Tapestry was also used to decorate furniture and clothing.

Early Chinese *k'o-ssu*

The earliest surviving examples of *k'o-ssu* date from the T'ang dynasty (AD 618–907). Eighth-century remains have been found in desert oases around Turfan in the Sinkiang Uighur Autonomous Region of China, and late T'ang fragments have been found in the Caves of the Thousand Buddhas (Ch'ien-fo-tung or Mo-kao-k'u) near the town of Tun-Huang in Kansu Province. It is thought that these weavings are probably not representative of the more fully developed *k'o-ssu* of the T'ang period because they show only simple repeating patterns of flowers, vines, ducks, lions, etc., and were found in relatively remote areas of Central Asia along the silk-trade route. In comparison is the more sophisticated 8th-century *k'o-ssu* that hangs in the Taima-dera, a temple near Nara, Japan. Based on the story of the T'ang dynasty priest Shan-tao, this 43-square-foot (four-square-metre) weaving is the oldest known complete Chinese wall tapestry.

During the Sung dynasty (960–1279) the imperial family encouraged painting and patronized the art of tapestry. An important weaving centre was at Ting-chou in Hopeh Province. Under the Yüan dynasty (1206–1368) a government factory for weaving *k'o-ssu* was established at Hangchow (Lin-an) in Chekiang Province. Characterized by their rich ornamental designs, the Hangchow *k'o-ssu* were frequently woven with gold and silver thread. Examples of tapestry from the Ming period (1368–1644) are rare and exquisite. The *k'o-ssu* executed during the K'ang-hsi (1661–1722), or rule of the great Manchu emperor of China, Hsüan-yeh, are the finest tapestries produced during the Ch'ing dynasty (1644–1911/12). They are distinguished for their delicate colouring and the use of philosophical and religious themes. Later Ch'ing *k'o-ssu* has survived in great abundance and shows a decided artistic and technical decline. This is especially evident in the frequent use of painting to perfect design details in 19th-century *k'o-ssu*.

Japanese *tsuzure*

The tapestry technique travelled from China to Japan in the late 15th or early 16th century during the Muromachi (Ashikaga) period (1338–1573). Japanese tapestry called *tsuzure-nishiki* (polychrome tapestry) differs from the Chinese *k'o-ssu* in its more pronounced surface relief. This is achieved through the use of thick cotton weft threads covered with silk, gold, or silver thread.

Paralleling the great period of sumptuous brocade manufacturing, the production of *tsuzure* flourished during the Tokugawa (Edo) period (1603–1867), especially in the early 17th century and throughout the entire 18th century (Figure 95). These polychrome tapestries were primarily used to decorate garments and for wrapping gifts; on rare occasions they were also used as wall hangings. Although the tapestry industry declined in quality in the 19th century, it has been revitalized in the 20th century. Monumental wall hangings and theatre curtains are woven in the textile factories of Ōsaka and Kyōto by both traditional Japanese and European tapestry techniques.

The history of the art in Korea remains obscure. Rather coarse wool tapestry-woven rugs with stylized motifs, however, are still produced there.

PRE-COLUMBIAN AMERICAS

The most skilled weaving in pre-Columbian America was achieved by the Andean Indian cultures of ancient Peru. The origins of tapestry weaving among these peoples are believed to date as early as the beginnings of the Christian Era. By the 6th and 7th centuries the technique of tapestry weaving was established, and a large number of pieces in this medium have survived, particularly from the 8th to the 12th centuries. Most of these tapestry weavings have been found in Peruvian coastal burial sites, where the dry desert climate prevented their deterioration. The dead were buried in clothes that display some of the most varied and skilled techniques of weaving and needlework ever current in any culture. Tapestry weaving was used principally to make garment decorations that were usually integral to the garment fabric. Narrow strips to ornament the edges of clothing were common, as were panels covering the entire surface of the *cuzma*, a poncho-like Indian shirt. Fragments of tapestry wall hangings have also survived.

According to chronicles written by Spanish colonizers and scenes painted on ancient Peruvian pottery, weaving was generally done by women whose great manual skill made up for the simplicity of the looms, which are still used by Indian craftsmen. The workmanship was extremely fine. Certain tapestry fragments have been found with 150 to 250 weft threads per square inch (60 to 100 per square centimetre). The warps of the tapestries are of undyed cotton, being, therefore, either white or brown. The wefts are of wool from the llama, guanaco, alpaca, or vicuña, with cotton sometimes used to obtain bright white. The tapestries are usually polychrome, for the range of available colours made with natural dyes was large. Strong colour contrasts were preferred to the use of subtly graded tones of colours, especially in the Inca period (*c.* 13th to 16th century). Compositions tended to bold conventionalized designs often of human or animal figures and elaborate geometric patterns. Plant motifs are comparatively rare.

By courtesy of the Museum of Fine Arts, Boston



Figure 96: Colonial tapestry, late 16th–17th centuries. In the Museum of Fine Arts, Boston. 2.29 × 2.12 m.

After the Spanish conquest, looms from Spain were imported by the viceroyalty of Peru, and the weaving of tapestry was continued during the colonial period. The skilled Inca and later mestizo weavers evolved a curious blending of European influences and Indian traditions.

Tapestry may also have been current in other developed pre-Columbian cultures of Central America and Mexico. Climatic conditions, however, have been destructive to textiles.

MIDDLE AGES IN EGYPT AND THE NEAR EAST

Tapestry weaving was done by the Copts, or Egyptian Christians, from the 3rd to about the 12th century AD. Their tapestries are of great interest not only because of their artistic quality and technical skill but also because

Coptic tapestry

they are a bridge between the art of the ancient world and the art of the Middle Ages in western Europe. Fragments from the 5th to the 7th century are particularly numerous, and the largest number of examples have survived in the Egyptian cemetery sites of Akhmim, Antinoë, and Saqqārah. As a result of a change in burial customs, perhaps attributable to Romanization and the widespread adoption of Christianity in Egypt, the ancient practice of mummification and its attendant ritual fell into disuse after the 4th century AD. The dead were subsequently buried in daily clothes or were wrapped in discarded wall hangings and tapestries. The clothing was ornamented with tapestry trimming, which was either woven into the fabric or attached to tunics and cloaks. Other burial furnishings included pillows and coverings. Tapestries were also used for the decoration of Christian churches, but few of these wall hangings have survived.

Coptic tapestries were woven with woollen wefts on linen warps, though a few with silk wefts have been preserved. Cotton wefts were occasionally used to obtain a brighter white. Primarily in the 7th century and perhaps also the 8th century, tapestry ornamentation was often supplemented by embroidery, as in border margins. In a special variant, which is not true tapestry, characteristic ornamental motifs such as meanders or other geometric repeats are executed with a free bobbin that follows the design without regard to consistency of weft direction.

Many of the early Coptic tapestries were done in a silhouette technique in which the motif or design was in a single dark colour, usually a tone of purple achieved by dying with madder and indigo, against a lighter background colour. After the 5th century, polychrome tapestries became increasingly common.

Many Coptic tapestry trimmings were woven with indigenous designs. Recurring motifs related to the ancient Egyptian funerary cult of Osiris and included the grape vine or ivy and the wine amphora. These motifs were considered appropriate to burial robes because of their relevance to revival in a life after death. Other favourite subjects were the hunter on horseback, boy-warriors, desert animals (especially the lion and the hare), creatures of mythology, dancing figures, and baskets of fruits and flowers (Figure 97). Christian subjects are as a rule late in date and are

By courtesy of (left) the Museum of Fine Arts, Boston, Ross Collection. (right) Yale University Art Gallery, the Hobart Moore Memorial Collection

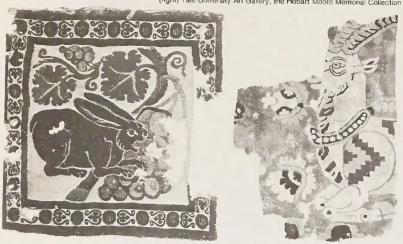


Figure 97: *Coptic and Persian tapestries.*
(Left) Coptic Medallion tapestry with rabbit and grapes, Egyptian, 4th–7th centuries AD. In the Museum of Fine Arts, Boston. 21.5 × 21.0 cm. (Right) Fragment of Persian dovetailed tapestry with ibex, probably Sāsānian period (*c.* AD 224–651). In Yale University Art Gallery. 35.74 × 27.30 cm.

mostly figures of saints, standing or on horseback, against a red background. Depictions of biblical stories are rare. Some of the Coptic designs were copied, in a more or less distorted manner, from those woven into silk textiles imported from Syria.

Islāmic tapestry

After the invasion of Egypt by the Muslims in 640, the quality of Coptic tapestry began to deteriorate, although the industry continued to flourish by adapting itself to the tastes of the conquerors. During the Tūlūnid period (868–905) bands of tapestry trimming in wool or often in silk, occasionally with metal-thread enrichments, were woven into white or dark green linen garments. In the Fātimid period (909–1171) silk tapestry weaving in golden yellow and scarlet became common. The motifs of the Islāmic period of Egyptian weaving were often interlacing geometric patterns frequently enclosing inscriptions or highly stylized small birds, animals, and flowers. Many of these inscriptions merely simulate writing, but many are legible. Giving religious phrases or the names and titles of rulers, they are in handsome angular Kufic scripts on earlier pieces and in cursive scripts later.

From the 6th to the 8th century AD, and doubtless from then on, striking wool tapestries were being made in Syria corresponding in style to the contemporary silk textiles with animals or birds in energetic heraldic stylization, framed in roundels, and almost always on a red ground. Later, from the 11th to the 13th century, highly distinctive silk- and gold-thread tapestries were produced in Syria incorporating pagan motifs from classical antiquity.

Fewer specimens of Persian tapestries have survived, but one notable fragment, now in the Moore Collection at Yale University, bears an ibex in the style of the Sāsānian period (Figure 97). A single piece from the Seljuq period (11th century) established a continuation of the use of the tapestry technique, which reappears in the 16th century (intermediate examples apparently having all been destroyed) as the medium for rich silk- and metal-thread rugs, of which only three are known still to exist (also in the Moore Collection, New Haven, Connecticut), though others are illustrated in Persian miniatures. The modern descendants of these are kilims, or pileless carpets woven by the tapestry technique. Common to the entire Near East, these rugs are especially produced in the Caucasus and Asia Minor, as well as in parts of eastern Europe. Occasionally silk, they are more often wool with simple geometric patterns in bold colours.

EARLY MIDDLE AGES IN WESTERN EUROPE

Numerous documents dating from as early as the end of the 8th century describe tapestries with figurative ornamentation decorating churches and monasteries in western Europe, but no examples remain, and the ambiguity of the terms used to refer to these hangings makes it impossible to be certain of the technique employed. The 11th-century so-called Bayeux Tapestry depicting the Norman Conquest of England, for example, is not a woven tapestry at all but is a crewel-embroidered hanging.

Like the art of stained glass, western European tapestry flourished largely from the beginnings of the Gothic period in the 13th century to the 20th century. Few pre-Gothic tapestries have survived. Perhaps the oldest preserved wall tapestry woven in medieval Europe is the hanging for the choir of the church of St. Gereon at Cologne in Germany. This seven-colour wool tapestry is generally thought to have been made in Cologne in the early 11th century. The medallions with bulls and griffons locked in combat were probably adapted from Byzantine or Syrian silk textiles. The "Cloth of Saint Gereon" is thematically ornamental, but an early series of three tapestries woven in the Rhineland for the Halberstadt Cathedral were narrative. Dating from the late 12th and early 13th centuries, these wool and linen hangings are highly stylized and schematic in their representations of figures and space, with all forms being outlined. The "Tapestry of the Angels," showing scenes from the life of Abraham and St. Michael the Archangel, and the "Tapestry of the Apostles," showing Christ surrounded by his 12 disciples, were both intended to be hung over the cathedral's choir stalls and therefore are long and narrow. The third hanging, called the "Tapestry of Charlemagne Among the Four Philosophers of Antiquity," is a vertical wall hanging related to works produced by the convent at Quedlinburg in the German Rhineland during the Romanesque period of the 12th and early 13th centuries.

Romanesque tapestry

Fragments of a tapestry with traces of human figures and trees reminiscent of hangings described in the Norse sagas were found in an early 9th-century burial ship excavated at Oseborg in Norway. One of the major works of Romanesque weaving is a more complete tapestry dating from around the end of the 12th or early 13th century



Figure 98: "April and May," fragment known as the "Baldishol Tapestry," after 1190. In the Kunstindustrimuseet, Oslo. 1.18 × 2.00 m.
By courtesy of the Kunstindustrimuseet, Oslo

that was made for the Norwegian church of Baldishol in the district of Hedemark. Originally a set of wool hangings on the 12 months of the year, only the panels of April and May have survived (Figure 98). The pronounced stylization of the images relates these tapestries to those executed for Halberstadt Cathedral.

14TH CENTURY

In the 14th century the western European tradition of tapestry became firmly established. At that time the most sophisticated centres of production were in Paris and Flanders. Large numbers of tapestries are recorded in inventories. The more luxurious standards of living being adopted by the wealthy of the Gothic period extended the use of tapestries beyond the customary wall hangings to covers for furniture. Survivals of 14th-century workmanship, however, are rare, and the most important of these were produced by Parisian weavers. The outstanding example of their art is the famous Angers Apocalypse (Musée des Tapisseries, Angers, France), which was begun in 1377 for the Duke of Anjou by Nicolas Bataille (flourished *c.* 1363–1400). This monumental set originally included seven tapestries, each measuring approximately 16.5 feet in height by 80 feet in length (5.03 by 24.38 metres). Based on cartoons drawn by Jean de Bandol of Bruges (flourished 1368–81), the official painter to Charles V, king of France, only 67 of the original 105 scenes have survived. A slightly later series (*c.* 1385) possibly woven in the same Parisian workshop is the "Nine Heroes" (Metropolitan Museum of Art, The Cloisters, New York). This set is not a religious narrative but illustrates the chivalric text *Histoire des Neuf Preux* ("Story of the Nine Heroes") by the early 14th-century wandering minstrel, or jongleur, Jacques de Longuyon.

Paris tapestry

Flanders, particularly the city of Arras, was the other great centre of the tapestry industry in 14th-century Europe. The tapestry produced there had such an international reputation that terms for tapestry in Italian (*arrazzo*) and Spanish (*drap de raz*) and English (arras) were derived from the name of this Flemish city. Long a medieval centre of textile weaving, Arras became an important tapestry centre when the leading citizens decided to create a luxury industry to alleviate the economic crisis caused by a decline in the sale of Arras textiles due to the popularity of cloth from the Flemish region of Brabant.

15TH CENTURY

Arras tapestry

The greatest tapestries of the 15th century were produced in the Flemish cities of Arras, Tournai, and Brussels. In the first half of the century it was Arras that particularly prospered under the patronage of the dukes of Burgundy. Duke Philip the Good (1396–1467) had a specially designed building erected in the city to allow for better conservation of his tapestry collection. Between 1423 and 1467 no fewer than 59 master tapestry weavers were working in Arras, but following the French siege of the city in 1477 under King Louis XI the industry declined. After approximately 1530 it was no longer active. While the importance of Arras waned, that of Tournai and eventually Brussels waxed—their tapestries becoming the most sought after in the late 15th century. Local identification marks did not become general until the 16th century, and continual intercourse between the various medieval centres of tapestry making, particularly Arras and Tournai, adds to the difficulty of determining where individual tapestries were made. Despite the prestige of Arras workmanship, it is ironic that only one set of tapestries dating from 1402 is inscribed with the actual name. Large fragments showing scenes from the lives of St. Piat and St. Eleutherius survive in the cathedral of Tournai, for which they were commissioned. The imagery of these tapestries, like that of most Gothic hangings, was closely related to the styles of painting current at the time. Other important examples of supposed Arras tapestries inspired by Franco-Flemish book miniatures or paintings on wood panels include the early 15th-century tapestry of "The Annunciation" (Metropolitan Museum of Art, New York), which was probably woven after a cartoon by Melchior Broederlam (active 1381–*c.* 1409), and the "Court Scenes" (Musée des Arts Décoratifs, Paris), related to the *Très Riches Heures du duc de Berry* illuminated by the brothers Limburg (active early 15th century).

Whether a tapestry is an Arras or not is usually determined by comparison with the "History of St. Piat and St. Eleuthère." One of the finest works so attributed is the early 14th-century fragment from the set in the Museo Civico at Padua, Italy, illustrating the *Geste of Jourdain de Blaye*, a medieval chivalric story adapted from the ancient Greco-Roman romance *Apollonius of Tyre*.

Tournai tapestry

The craft, practiced since the end of the 13th century at Tournai, proved so prosperous that in 1398 a regulation concerning production was published. It is the oldest known ordinance regulating the craft of tapestry weaving. Among partially surviving tapestries ordered in the late 15th century by the Court of Burgundy were two sets produced by the weaver and tapestry merchant Pasquier Grenier (died 1493) for Philip the Good. One set, "The Story of Alexander" (Galleria Doria-Pamphili, Rome), was purchased in 1459, and the other, "The Knight of the Swan" (St. Katherine, Kraków, Poland, and Österreichisches Museum für Angewandte Kunst, Vienna), was bought in 1462.

Cited by many scholars as an example of mid-15th-century Tournai weaving under the influence of Arras are the four renowned tapestries of "The Hunts of the Dukes of Devonshire" (Victoria and Albert Museum, London). Typical of the developed late Gothic Tournai style are the compacted vertical compositions of "The Story of Strong King Clovis" (mid-15th century; Musée de l'Oeuvre Notre-Dame, Reims, France) and "The Story of Caesar" (*c.* 1465–70; Historisches Museum, Bern, Switzerland). Many of the attributed Tournai weavings are heavily outlined and have a solemnity that contrasts to the more fanciful nature of Arras weavings. A sense of monumentality is created by the immense size of many of these supposed Tournai weavings and by the way the vast surfaces are densely filled with superimposed imagery.

A producer of tapestry since the 14th century, in the 15th century Brussels vied with Arras and Tournai. By mid-century, Brussels was noted for its highly skilled repro-

Giraudon—Art Resource



Figure 99: "The Adoration of the Magi," Brussels altarpiece tapestry, 1466–88. In the Cathedral of Sens, France. 1.38 × 3.31 m.

ductions of religious paintings by Flemish masters of late Gothic realism, such as in the tapestry of "The Adoration of the Magi" (Figure 99). These panels were called "altarpiece tapestries" because they were usually intended for churches or private chapels, where they either were used as an altar cloth or antependium or were hung behind the altar as an altarpiece or fabric retable. In scale, altarpiece tapestries approximated the dimensions of the painting they copied and were, therefore, much smaller in size than the muralesque wall hangings of Arras and Tournai. Silk was commonly used to obtain the greater degree of naturalistic detail essential in reproducing a painting.

Tapis d'or and millefleurs tapestries

In the late 15th and early 16th centuries, Brussels also became famous for its production of *tapis d'or,* or "golden carpets," so called because of the profuse use of gold threads. Examples such as "The Triumph of Christ," popularly known as the "Mazarin Tapestry" (*c.* 1500; National Gallery of Art, Washington, D.C.), are characterized by their richness of effect.

Perhaps the best known late Gothic hangings were the fanciful tapestries usually referred to as millefleurs ("thousand flowers"). A red or dark-blue ground strewn with flora and fauna sometimes serves as a setting for heraldic devices such as in the late 15th-century tapestry with the coat of arms of Philip the Good (Historisches Museum, Bern, Switzerland) or acts as a background for scenes of the chivalric aristocratic life during the late Middle Ages, such as in "The Hunt of the Unicorn" (Metropolitan Museum of Art, The Cloisters, New York) or "The Lady with the Unicorn" (Musée de Cluny, Paris). The origin of millefleurs tapestries is disputed, but it is thought that they were woven in the Flemish workshops of Brussels and Bruges and by itinerant weavers in the Loire Valley of France.

Itinerant Flemish and French weavers, setting up their looms in cities where there was temporary employment, carried tapestry weaving to Italy as early as the 15th century. Before the 16th century, however, most tapestries were bought in France and Flanders. Small workshops attached to the courts of various Italian nobles sporadically appeared for brief periods in Siena, Brescia, Todi, Perugia, Urbino, Mantua, Modena, Genoa, and Ferrara. The only one of importance was the Flemish-directed workshop of Ferrara, established around 1445 by the duke Lionello d'Este, who commissioned the famous Ferrarese early Renaissance painter Cosmè Tura (*c.* 1430–95) to make cartoons for his weavers.

16TH CENTURY

Two new trends became apparent in the 16th century. The first, brought about by war and persecution in Flanders, resulted in the widespread diffusion of the Flemish art of tapestry weaving. Many Flemish artisans in the 16th century were forced to become refugees. Some grouped together to live the life of travelling craftsmen, while others attempted to reestablish their trade abroad. Flemish weavers were welcomed everywhere as carriers of a great tradition. Such itinerant masters established shops from England to Italy. The second important new trend emanated from Italy and reflected the superiority attached by the Italian Renaissance to the art of painting. The decisive step, which was to bring about the subordination of weaving to painting for more than 400 years in the art of tapestry, was taken when Pope Leo X commissioned the famed weaver Pieter van Aelst (flourished late 15th–early 16th century) of Brussels to make a series of tapestries illustrating the "Acts of the Apostles" from cartoons produced between 1514 and 1516 by Raphael (1483–1520). Little or no concession had been made to the tapestry medium for which the cartoons were intended, but the tapestries were a great success, and numerous copies of them were subsequently made.

The ascendancy of Brussels tapestry

The occupation of Arras by the French in the late 15th century and successive sieges of Tournai in the early 16th century contributed to the rise of Brussels as the leading tapestry centre of Flanders—a position it maintained until the 17th century. The patronage of the papacy and the imperial houses of Spain and Austria, along with other European royalty and the skill of its weavers, who were among the finest in Europe, combined to establish the international reputation of Brussels tapestry. The industry was controlled by a monopoly of rich merchants. Tapestry making proved so prosperous in the period between 1510 and the outbreak of the Peasants' War in 1568 that the industry had to be protected by regulations against frauds and forgeries. A number of communal ordinances followed one another in rapid succession, the most important being that of 1528, requiring each tapestry woven in Brussels to bear the mark of the city—a flat red shield flanked by two *B*'s standing for Brussels and the province of Brabant. The same imperial edict issued by Emperor Charles V also required manufacturers and merchants to use the signature or monogram of the master weaver or workshop.

Work of van Orley

It is the designs of the Flemish painter Bernard van Orley (1492?–1541) that are most characteristic of the Renaissance style of Brussels tapestry. Van Orley attempted to reconcile the traditions of late Gothic northern realism and the monumentality and idealism of Italian Renaissance art with the artistic potential of the tapestry medium. His earlier works, such as "The Legend of Our Lady of Le Sablon" (Musées Royaux d'Art et d'Histoire, Brussels) and "The Revelation of St. John" (1520–30; Patrimonio Nacional, Spain), still show compositional elements that link them to medieval Flemish art. Later, his work was influenced by the cartoons of Italian artists that were woven in Brussels workshops, such as Raphael's "Act of the Apostles" and the designs for "The Story of Scipio" and "Fructus Belli," executed by Raphael's disciple, the Mannerist painter and architect Giulio Romano (1499–1546). Van Orley adapted the Italians' preference for monumentality and their feeling for depth and sculptural modelling to Flemish tastes and traditions for genre and naturalistic detail in sets such as "The Battle of Pavia" (Museo e Gallerie Nazionali di Capodimonte, Naples), "The Story of Abraham" (Madrid and Vienna), "The Story of Tobias" (Vienna), and "The Hunts of the Emperor Maximilian I" (before 1528; Louvre, Paris). Among his followers in the first half of the 16th century were the Flemish painters Pieter Coecke van Aelst (1502–50), Jan Vermeyen (*c.* 1500–59), and Michel Coxcie (1499–1592). It was not only the cartoonists of Brussels who achieved international reputations but also the weavers of the early 15th century. Among the best known are Pieter van Aelst, Pieter and Willem Pannemaker, and Frans (active *c.* 1540–90) and Jacob Geubels (active *c.* 1580–1605).

Other limited centres of tapestry making in 16th-century Flanders were Antwerp, Bruges, Enghien, Oudenaarde, Grammont, Alost, Lille, and Tournai. Perhaps the most distinctive type of tapestry produced in these cities was the *verdures* of Enghien and Oudenaarde. French tapestry weaving, after its eclipse in the 15th century when nomadic weavers seem to have been more active than established shops, owes much of its eventual prestige to an unusual degree of royal patronage. This resulted in the 17th century in the foundation of the Gobelins and Beauvais state factories, the names of which have now become household words. A prelude to this development was the factory established by Francis I in 1538 near Paris

By courtesy of the Österreichische Nationalbibliothek, Vienna

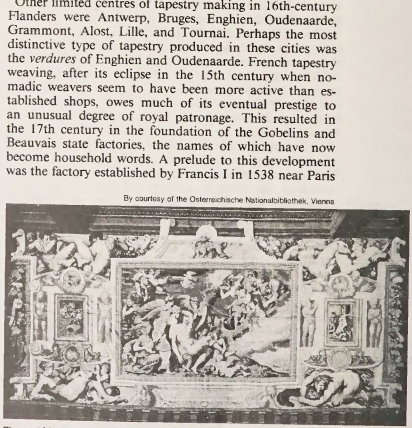


Figure 100: "The Death of Adonis" (after a mural in the Galerie des Réformes, Fontainebleau), tapestry by Francesco Primaticcio, 1541–50. In the Kunsthistorisches Museum, Vienna. 3.30 × 6.40 m.

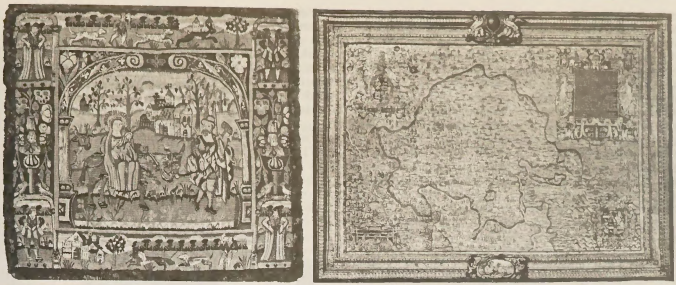


Figure 101: *English 16th- and 17th-century tapestries.*
(Left) Tapestry cover with "The Flight into Egypt," *c.* 1600. In the Victoria and Albert Museum, London. 28.26 × 20.32 cm. (Right) Topographical tapestry depicting "Warwickshire," from late-16th-century designs by William Sheldon. In the Warwick Castle Museum, Warwick, England. 3.99 × 5.26 m.
By courtesy of (left) the Victoria and Albert Museum, London, (right) the County of Warwick Museum, England

at the château of Fontainebleau to make tapestries for his royal residences. Staffed by Flemish weavers, the cartoons were largely furnished by two Italian Mannerist artists, Francesco Primaticcio (1504–70) and Rosso Fiorentino (1494–1540), who were court painters to the King. The six tapestries (Kunsthistorisches Museum, Vienna), based on their murals for the Galerie des Réformes in the château, are the first tapestries in which sculpture as well as painting is imitated in the highly illusionistic manner of a trompe-l'oeil (fool-the-eye) effect (Figure 100).

French trompe-l'oeil tapestries

The Fontainebleau workshop, which was active for only 12 years, provided the springboard for subsequent developments in Paris, where in 1551 Henry II established and endowed with special privileges the Hôpital de la Trinité factory.

In the first third of the 16th century, Franco-Flemish weavers and small court workshops continued to supply the only indigenous Italian tapestry. Weaving was done in Genoa, Verona, Venice, Milan, and Mantua. The first internationally important Italian tapestry factory was established in 1536 in Ferrara by Duke Ercole II of the House of Este. The Arrazeria Medicea founded in 1546 in Florence by the Medici grand duke Cosimo I (1519–74) was the most important tapestry factory instituted in Italy during the 16th century and survived into the early 18th century. It was headed initially by the famous mid-15th-century Flemish weavers Nicolas Karcher and Jan van der Roost, both of whom had worked in the Ferrara workshop of Duke Ercole II.

Cartoons were designed by such leading Mannerist artists of Florence as Jacopo Pontormo (1494–1556/57), Francesco Salviati (1510–63), Il Bronzino (1503–72), and Bachiacca (1494–1557), who designed the "Grotesques" (*c.* 1550; Uffizi, Florence), one of the most famous and influential tapestry sets produced by the Arrazeria Medicea.

The major textile art in medieval England was embroidery. When woven tapestries were needed, they were imported from Flanders. Although occasional references to Arras weavers in England date from the 13th century and a few indigenous armorial tapestries have survived from the 15th century, it was only after the middle of the 16th century that the English organized tapestry works. The first important workshops were set up in Bercheston (Warwickshire) by a wealthy squire, William Sheldon (died 1570). They initially produced cushion covers and small hangings of heraldic and ornamental subjects. A later specialty of these shops was a series of topographic tapestries. Woven in 1588 from contemporary maps of the Midland counties, these tapestries featured bird's-eye views of hills, trees, and towns, surrounded, according to the custom of the period, by Flemish-styled borders of architectural and figural ornament (Figure 101, right). Many of the men who worked in these shops were Flemings who had fled the mid-16th-century religious persecutions in the Lowlands.

German court factories

Germany was one of the first regions to receive Flemish weavers fleeing religious persecution in the Lowlands. Their small workshops prospered in such cities as Cologne, Hamburg, Kassel, Leipzig, Torgau, Lüneburg, Frankenthal, and Stuttgart. Most of the works produced were in the Flemish style. In Switzerland, on the other hand, where tapestry making had flourished in the 14th and 15th centuries, the industry almost ceased to exist except around Basel and Lucerne.

17TH AND 18TH CENTURIES

It was due to the initiative of Henry IV, whose planning of his nation's economy emphasized the luxury production that has since been commercially important in France, that decisive steps were taken in establishing a French tapestry industry. In 1608 Henry gave official recognition to the French workshop (using the high-warp method) of Girard Laurent and Dubout by establishing them in the Louvre, and at the same time he encouraged the immigration of Flemish weavers practicing the low-warp method who would help Paris to compete with the flourishing industries of Brussels and Antwerp.

At the turn of the 16th–17th centuries, two Flemish weavers had been taken to France by government arrangement to establish low-warp looms in Paris: François de La Planche (or Franz van den Planken; 1573–1627) and Marc de Comans (1563–before 1650). Satisfactory working conditions were found for them in the old Gobelins family dyeworks on the outskirts of the city, and so began the establishment commonly known by that name that has lasted ever since. One of its first ambitious productions was an allegorical invention lauding Catherine de Médicis under the guise of Artemisia. The cartoons for this set were chiefly by the French Mannerist painter Antoine Caron (*c.* 1515–93). The Baroque verve and vitality of the Flemish painter Peter Paul Rubens (1577–1640) and Simon Vouet (1590–1649) brought new life to French designs in the early 17th century.

De La Planche died in 1627 and was succeeded by his son, who broke with the Comans family and moved to the Faubourg Saint-German-des-Près, leaving the Comans at the Gobelins. Competition became bitter, but both continued to produce a considerable quantity, as well as good quality, until they were superseded in 1662 by the royal factory, which purchased the Gobelins works at its location.

Establishment of the Gobelins factory

The Gobelins was officially established in 1667, receiving the title Manufacture Royale des Meubles de la Couronne ("Royal Factory of Furnishings to the Crown"). Initially it included all the king's artisan corps (tapestry weavers, cabinetmakers, goldsmiths and silversmiths, etc.) that produced furnishings for the royal residences, especially the

château of Versailles. Louis XIV's finance minister, Jean-Baptiste Colbert (1619–83), always alert to profitable opportunities, recruited skilled personnel not only from the de La Planche and Comans shops but also from the old Louvre enterprise and thus established a new tapestry works with both high- and low-warp looms. The Gobelins' first director was the painter Charles Le Brun (1619–90), who had managed the short-lived royal tapestry works established in 1658 by Colbert's predecessor, Nicolas Fouquet (1615–80), at his château of Vaux-le-Vicomte near Paris. Le Brun applied himself with prodigious energy to his new position and proved to have a special talent for the task of celebrating the glory of Louis XIV. Among the most important sets he designed were "The Elements," "The Seasons," "The Child Gardeners," "The Story of Alexander," and, above all, the "Life of Louis XIV" and the "Royal Residences" (most of these sets are in the possession of Mobilier National in Paris).

When Le Brun died, the painter Pierre Mignard (1612–95) became director. The draining of the royal treasury closed the Gobelins in 1694. The factory opened again in 1699, when a lighter spirit was introduced into tapestry design by the decorative inventions, especially grotesques, of Claude Audran III (1658–1734), who designed such sets as "The Grotesque Months" and "The Portières of the Gods." Louis XV (1710–74), in his turn, was celebrated in a set of "Hunts" by the Rococo painter Jean-Baptiste Oudry (1686–1755). Oudry was director of the Gobelins from 1733 until his death in 1755, when he was succeeded by François Boucher (1703–70), the outstanding artist-director of the 18th century. Boucher and Charles-Antoine Coypel (1694–1752), a Rococo painter, designed many of the popular *alentours* tapestries, in which the central subject, presented as a painting bordered by a frame simulating gilded wood, is eclipsed by the rich use of ornamental devices surrounding it. Boucher's "Loves of the Gods" were also *alentours* and enjoyed a great success and popularity, especially among the English nobility. "The Story of Don Quixote" (Mobilier National, Paris) was designed by Coypel and woven nine times between 1714 and 1794.

Alentours tapestry

Oudry's sophistication and polished elegance posed new problems for the weavers. Now indeed it was necessary for them to learn to paint with a bobbin, and to this end hundreds of new dyes were perfected for both wool and silk, until about 10,000 hues were available, to effect almost imperceptible tonal modulations; and interlocking of the wefts was introduced to render the transitions practically invisible, while the finest textures practical were used.

The Gobelins succeeded in surviving the French Revolution. Napoleon as emperor, like Louis XIV, desired an art of apotheosis and ordered a set of tapestries (1809–15; Mobilier National, Paris) that were devoted to his reign. Paintings by such French Neoclassical painters as Jacques-Louis David (1748–1825), Carle Vernet (1758–1836), and Anne-Louis Girodet-Trioson (1767–1824) were woven into tapestries in the late 18th and early 19th centuries.

The Beauvais factory

Another major state-subsidized factory established in 1664 at Beauvais had been carried on by two Flemings, Louis Hinart for 20 years and Philippe Behagle for 27 more. It was administered in much the same way as the Gobelins. Beauvais, however, was a private enterprise with royal patronage intended to produce tapestries for the nobility and the rich bourgeoisie, while Gobelins' work was only for the king.

Two types of decorative panels were particularly developed at Beauvais in the late 17th century, the architectural composition and the grotesque. The former, such as in the set of "Marine Triumphs" (1690; Banque de France, Paris), usually shows a complex fantasy architecture reminiscent of Baroque stage sets. In the latter, architectural tracery defines a complex of panels, framing a medley of festoons, scarves, vases, musical instruments, putti, masks, and comedy actors, such as in "The Rope Dancer and the Dromedary" (*c.* 1689; Mobilier National, Paris).

Both Oudry and Boucher designed for the Beauvais factory. The "Fables of La Fontaine," by Oudry, were among the most popular tapestries of the 18th century. In 1736 Boucher designed Italian genre scenes for the set "Village Festivities" and later in the "Second Chinese Set" did Chinese fantasies. He also designed various pastoral scenes with titillating overtones. The Beauvais factory became noted for tapestry to upholster furniture with and panels for screens. These were usually floral designs and in the 19th century were especially fashionable in finely woven silk (Figure 102). By the end of the century, though technical standards were maintained, artistic deterioration set in.

By courtesy of the Mobilier National, Paris; photograph, Visages de France



Figure 102: Fauteuil (armchair) covered with Beauvais silk tapestry, first quarter of the 19th century. In the Mobilier National, Paris.

The Aubusson–Felletin factories

Factories at the neighbouring old tapestry-making communities of Aubusson and Felletin, which had operated for a century and a half as modest private undertakings, were allowed to use the royal Aubusson mark as of 1665. From a small house industry, in which weavers independently produced inexpensive tapestries on their own low-warp looms for a bourgeois clientele, the tapestry makers soon produced hangings, upholstery fabrics, and carpets in Aubusson. The most effective tapestries are the chinoiseries, or genre fantasies set in China, a theme popular in Rococo art. Those designed by Jean Pillement (1728–1808) are especially famous (Figure 103). Coarse and rather dull, the *verdures*, or "garden tapestries," which were the first Beauvais tapestries, were made in quantities. Aubusson architectural panels either imitate those of the Gobelins and Beauvais factories, often with more complex elements and the addition of animals, or depict a damasked wall hung with a painting or cluster of decorative objects and garlands. The factory was especially successful in its production of carpets with conventional geometric ornamental motifs or floral designs.

By courtesy of the Mobilier National, Paris; photograph, Alo Photo



Figure 103: *18th-century French tapestry.* "The Pagoda," chinoiserie tapestry woven at Aubusson and designed by Jean Pillement, 18th century. In the Mobilier National, Paris. 2.85 × 5.84 m.

The dominant influence on the Brussels industry of the 17th century was the Antwerp painter Peter Paul Rubens, whose most famous set was the "Triumph of the Eucharist" (1627–28). Imitations and adaptations of his style were legion. Heavy and elaborate columns were often substituted for side borders. On a more modest scale are the tapestry versions of genre paintings by David Teniers the Younger (1610–90), in which the border frequently simulated the actual picture frame.

The first major tapestry factory to be established in Germany was founded in 1604 in Munich by Duke Maximilian of Bavaria. The designers and weavers were all Flemish. Although the factory closed after only 11 years of operation, the quality of its workmanship was outstanding. Following the loss of religious freedom in France when the Edict of Nantes was revoked in 1685, many French weavers, especially from the Aubusson factory, sought refuge from persecution in Germany as had the persecuted Flemish weavers of the 16th century. The workshop established in 1686 in Berlin by the great elector Frederick William of Brandenburg (1620–88) employed many of these displaced Aubusson weavers. It produced tapestries mainly for the palaces built by the Great Elector's son, King Frederick I of Prussia (1657–1713), after whose death the factory closed.

French designers and weavers continued to produce a large number of tapestries in the 18th century. Tapestry production was centred principally in Munich, Berlin, Würzburg, Dresden, Schwabach, and Erlangen.

Scandinavian folk tapestry

In Scandinavia tapestries for the Danish and Swedish royalty were woven in Copenhagen and Stockholm. The weavers and designers were usually French and Flemish. Norway and Sweden continued to produce folk tapestries. Of the nearly 1,300 registered Norwegian tapestries, approximately 1,250 originated in small rural communities. These tapestries were usually coarse in texture, stylized and schematic in design, and boldly coloured (Figure 104).

James I established in 1619 by royal charter a factory of tapestry weaving at Mortlake near London. It was staffed by 50 Flemings. Philip de Maecht, a member of the famous late 16th- and 17th-century family of Dutch tapestry weavers, was brought from the de La Planche-Comans factory in Paris, where he had been the master weaver, to hold the same position at Mortlake. The royal factory flourished under the patronage of the Stuart monarchs James I and Charles I. Many of the early tapestries produced at Mortlake were modeled after hangings woven in Brussels. Rubens supplied cartoons and in 1623 suggested to Charles I the purchase of seven of the Raphael cartoons for the "Acts of the Apostles." A new set was woven from these cartoons at Mortlake and is preserved at the Mobilier National in Paris. The redesigned borders have been attributed to the renowned Flemish painter to the English court, Sir Anthony Van Dyck (1599–1641). Although the factory weathered the Puritan austerity of the Commonwealth period, it deteriorated under Charles II and closed in 1703.

From the late 17th century Francis Poyntz (died 1685) and his brothers had a studio in Soho, where a number of weavers originally employed in the royal factory produced a distinct style of tapestry based on Chinese and Indian lacquerwork.

Italian tapestry factories

Cardinal Francesco Barberini, the nephew of Pope Urban VIII, in 1633 established a tapestry factory in Rome. Even though it enjoyed papal patronage, it lasted only until 1679. Clement XI tried to establish another Roman tapestry works in 1710, which also failed. During the 18th century other small factories briefly existed in Turin and Naples. They were staffed mainly with weavers left unemployed by the closing of the Medici factory (Arrazeria Medicea) in Florence.

During the 15th and 16th centuries Franco-Flemish tapestries were imported in great quantities, and Flemish weavers were invited to Spain in order to repair and care for them. For a short time in the 17th century a factory, established by Philip IV (1605–65), operated at Pastrana near Madrid. It was not until Philip V (1683–1746) established the Real Fábrica de Tapices y Alfombras de Santa Barbara (Royal Factory of Tapestries and Rugs of St. Barbara) in 1720 at Madrid, however, that important tapestry was produced in Spain. Initially, the weavers and director were Flemings. The first tapestries made at Santa Barbara were woven from the cartoons of such Flemish Baroque painters as David Teniers the Younger (1610–90) and Philips Wouwerman (1619–68) or based on famous paintings by such Italian artists as Raphael and Guido Reni (1575–1642). When the early Neoclassical painter Anton Raphael Mengs (1728–79) became director, the factory entered its most brilliant period of production. The Spanish painter Francisco Bayeu (1734–95) and his painter son-in-law Francisco Goya (1746–1828) were commissioned to make cartoons. From 1777 to 1790 Goya made 43 cartoons for the "Los Tapices" ("The Tapestries") series depicting Spanish daily life. The painted models for this are among the finest works of Goya's Rococo style.

The French destroyed the factory in 1808, but after the Napoleonic occupation, production was resumed until 1835. The tapestries produced during this period were largely copies of works woven in the 18th century.

A tapestry factory staffed by weavers from the Gobelins was established at St. Petersburg in 1716 by Tsar Peter the Great (1672–1725). Although tapestries were produced until 1859, production was often plagued with difficulties. The most striking designs were a set of grotesques (1733–38) and a series of portraits, of which those of Catherine the Great (1729–96) are the most noteworthy.

By courtesy of the Kunstindustrimuseet, Oslo



Figure 104: "The Feast of Herod," tapestry from Gudbrandsdal, Norway, 17th century. In the Kunstindustrimuseet, Oslo. 1.96 × 1.37 m.

19TH AND 20TH CENTURIES

The influence of the Industrial Revolution

Most 19th-century tapestries reproduced paintings or previously woven designs. The influence of the Industrial Revolution was inescapable, of course, not only in tools, materials, and dyes but in the new middle-class market and its demands. Machine-made tapestry, although an achievement in mechanical weaving, became a threat to the survival of the original handicraft.

The necessity for the revitalization and purification of the tapestry art was first recognized by the artists associated with the Arts and Crafts Movement in late 19th-century England. Decrying the loss of individual creativity, they revived the ideals of medieval craftsmanship in an attempt to counter the effects of industrialization on the decorative or applied arts. The leader and most im-



Figure 105: *Late 19th-century tapestries.*
(Left) "Angell Laudantes," tapestry designed by Edward Burne-Jones (workshop of William Morris, England), 1894. In the Victoria and Albert Museum, London. 2.25 × 2.00 m. (Right) "Swans," tapestry after Otto Eckmann (Scherrebek workshop, Germany), 1897. In the Museum für Kunst und Gewerbe, Hamburg. 2.35 × 7.65 m.
By courtesy of (left) the Victoria and Albert Museum, London, (right) the Museum für Kunst und Gewerbe, Hamburg

portant figure of the movement was the artist William Morris (1834–96), who established a tapestry factory at Merton Abbey in Surrey near London. For about 15 years he and his associates had been designing not only for looms but also for pictorial wall decorations and stained-glass windows. They were well prepared professionally, therefore, to design tapestries. Morris and the painter-illustrator Walter Crane (1845–1915) contributed cartoon sketches, but most Merton tapestries were designed by the Pre-Raphaelite painter Sir Edward Burne-Jones (1833–98; Figure 105, left). More venturesome than any of the Merton Abbey products were the tapestry designs made in the 1880s by the artist and architect Arthur Heygate Mackmurdo (1851–1942), who in 1882 founded the Century Guild, the first of many groups of artists-craftsmen-designers to follow the teachings of William Morris. This tradition, influenced by the tapestry revival in mid-20th-century France, has continued in Scotland. The most ambitious 20th-century tapestry designed by a British artist, Graham Sutherland's (1903–80) enormous "Christ of the Apocalypse" (1962) for Coventry Cathedral, was, however, woven on looms in Felletin, France. This is the largest tapestry ever to have been made there (78 feet 1 inch by 38 feet 1 inch; 23.8 by 11.6 metres).

Folk traditions in modern European tapestry

In Europe during the late 19th century there was a resurgence of tapestry based on folk traditions. This trend was already apparent in Norway shortly after 1890, when special efforts were made to base a modern tapestry art on native medieval weavings. The leaders were Gerhard Munthe (1849–1929), a well-known painter, and Frida Hansen (1855–1931), a weaver who studied the peasant craftsmanship of Norway and evolved an individual, light, and open weave. Somewhat later developments in Scandinavia occurred in Sweden and Finland. Märta Måås-Fjetterström (1873–1941) became the best known Swedish tapestry artist, and her atelier continued to produce excellent works. In Finland a freer, more colourful art, more delicately scaled, has been practiced by many; among the best known are Martta Taipale, Laila Karttunen, and, for damask tapestry, Dora Jung. In Norway and to a lesser degree in Denmark, similar work has been done. The church in the Scandinavian countries has been unusually receptive to this art. Traditional folk weaving was also behind the revival of tapestry making in several other countries after World War I, including Czechoslovakia and Hungary. Poland produced especially original designs executed in a remarkably free technique. Following the tradition of heavy-grained native weaving, mid-20th-century Polish designer-weavers such as Magdalena Abakanowicz (Figure 106, top right) and Wojciech Sadley used unconventional materials such as jute, sisal, horsehair, and raffia in abstract tapestries that emphasize the nature of the material, tactile stimulation, plasticity, or surface relief.

Germany, emulating Scandinavia, also began a revival of tapestry weaving around the turn of the century. In the state of Schleswig-Holstein a small tapestry industry was set up from 1896 to 1903 at Scherrebek (Figure 105, right), followed by similar enterprises at nearby Kiel and Meldorf. The most significant development, however, occurred at the design school of the Bauhaus, where tapestry was created during the 1920s and early 1930s. Abstract in composition, the Bauhaus designs were deeply rooted in the theory that the technology of the craft should be revealed in the work and in expressing the nature of the materials used, especially by the exploitation of heavy fibres as strong textural elements. Anni Albers, wife of the painter and Bauhaus instructor Josef Albers, became the chief practitioner of this kind of tapestry (Figure 106, top left). Like most modern tapestry weavers, she also designed for the textile industry. After World War II, tapestry works were established in Munich and Nürnberg, and individual weavers worked throughout Germany and in Vienna. Among the Germans, unlike the French, stained glass rather than tapestry generated greater enthusiasm as a revived craft in the post-World War II period. A few indi-

Bauhaus tapestry

By courtesy of (top left) Anni Albers, (bottom) the Mobilier National, Paris; photograph, (top right) Edita S.A., Lausanne, Switzerland, (bottom) Visages de France

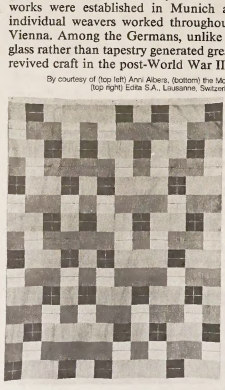
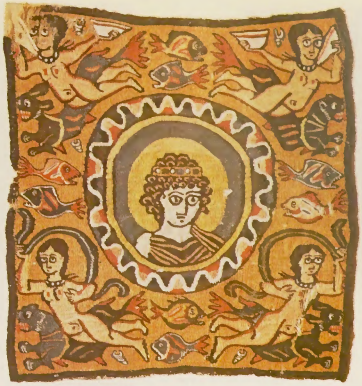




Figure 106: *20th-century tapestries.*
(Top left) Tapestry by Anni Albers, 1927. In the collection of the artist. 1.19 × 152 m. (Top right) "White Composition," tapestry after Madgalena Abakanowicz, 1962. In the collection of the artist. 1.05 × 1.80 m. (Bottom) "Mont Saint-Michel," after Henri-Georges Adam, 1965. In the Mobilier National, Paris, 4.02 × 5.65 m.



(Above) "St. Michael," detail from "Abraham and the Archangel Michael," Lower Saxony, mid-12th century. In Halberstadt Cathedral, East Germany. 1.10 × 10.26 m (whole tapestry).



(Right) "Square with Nereids." Coptic, Egypt, 6th century. In the Louvre, Paris. 30 × 30 cm.



"Two Ducks," detail from the *k'o-ssu* panel "Feng-huang in a Rock Garden." Ming period, China (1368–1644). In the Metropolitan Museum of Art, New York City. 2.20 × 1.75 m (whole tapestry).

Early tapestries of East and West



Pre-Columbian fragment from the coast of Peru, Late Coastal Tiahuanaco period (1000–1300). In the pre-Columbian collection of Dumbarton Oaks, Washington, D.C. 34.4 × 16.5 cm.

Plate 1: By courtesy of (top left) the Institut für Denkmalpflege, Halle, East Germany, (centre) the Metropolitan Museum of Art, New York, Seymour Fund, 1960, (bottom) Dumbarton Oaks, Washington, D.C., photograph, (top right) Photo Hachette



Detail from "The Bear Hunt," one of the Devonshire hunting tapestries series, Brussels, second quarter of the 15th century. In the Victoria and Albert Museum, London. 4.90 × 10.15 m (whole tapestry).

Tapestries of the 14th and 15th centuries



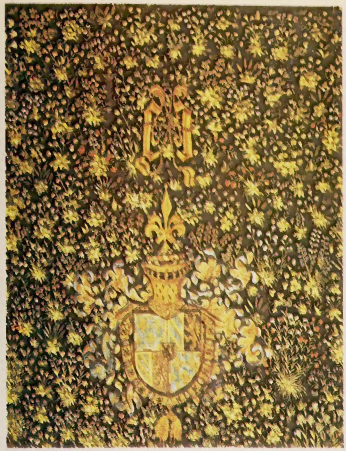


(Above) "St. Michael and the Dragon," from the set of seven tapestries in the "Angers Apocalypse," woven by Nicolas Bataille after the cartoons of Jean de Bandol, second half of the 14th century. In the Musée des Tapisseries, Angers, France. 2 m × 2 m 60 cm.

(Left) "The Annunciation," school of Arras, Flemish, early 15th century. In the Metropolitan Museum of Art, New York City. 3.46 × 2.90 m.

Plate 2: By courtesy of (bottom left) the Metropolitan Museum of Art, New York, gift of Harriet Barnes Pratt, 1949, in memory of her husband, Harold Irving Pratt; photographs, (top) J.R. Freeman & Co. Ltd., (bottom right) Photo Hachette

Plate 3: By courtesy of (top left) the Bernisches Historisches Museum, Switzerland, (bottom) the National Gallery of Art, Washington, D.C., Widener Collection; photograph, (top right) Lauros—Giraudon Art Resource



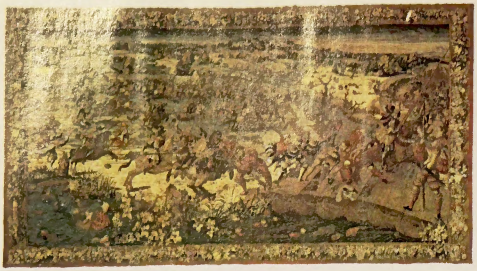


(Above) "The Lady with the Unicorn," one of the six pieces of the tapestry, Loire workshop, late 15th century. In the Musée de Cluny, Paris. 3 × 3.10 m.

(Left) "Coat of Arms of Charles the Bold," Tournai, 1450–70. In the Bernisches Historisches Museum, Bern. 3.05 × 6.86 m (whole tapestry).



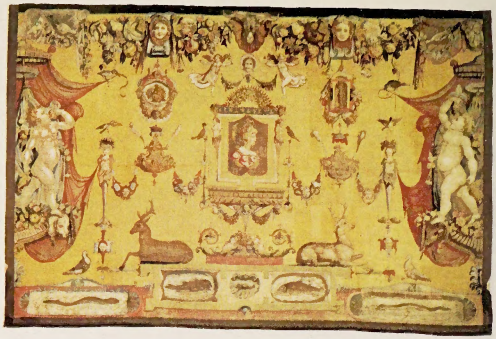
"The Triumph of Christ," known as the "Mazarin Tapestry," Brussels, *c.* 1500. In the National Gallery of Art, Washington, D.C. 3.30 × 3.90 m.



(Left) "Capture of Francis I," one of the panels of "The Battle of Pavia," Flemish workshop after cartoons by Bernard van Orley (1492?–1542). In the Museo e Galleria Nazionali di Capodimonte, Naples. 4.35 × 7.89 m.

(Below) Panel with grotesques woven by the workshop of N. Karcher after a cartoon by Bachiacca, *c.* 1550. In the Uffizi, Florence. 2.20 × 4 m.

Tapestries of the 16th, 17th, and 18th centuries





"The Miraculous Draught of Fishes," woven by Pierre d'Alost after a cartoon by Raphael, 1516–19. In the Vatican Museums. 4.95 × 4.39 m.



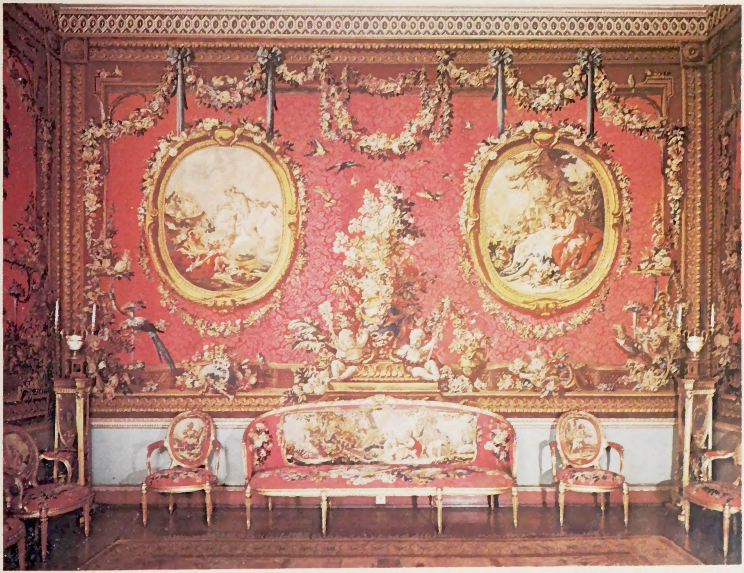
"Hercules and One of Diomedes' Stallions," from the set of four tapestries in "The Labours of Hercules," Oudenaarde, Belgium, second half of the 16th century. In the Louvre, Paris. 3.55 × 4.44 m.

Plate 4: (All but bottom right) SCALA—Art Resource. (bottom right) Photo Hachette
Plate 5: By courtesy of (bottom) Osterley Park, London; photographs, (top left) SCALA—Art Resource, (top right) J.E. Bulloz, (bottom) J.R. Freeman & Co. Ltd.

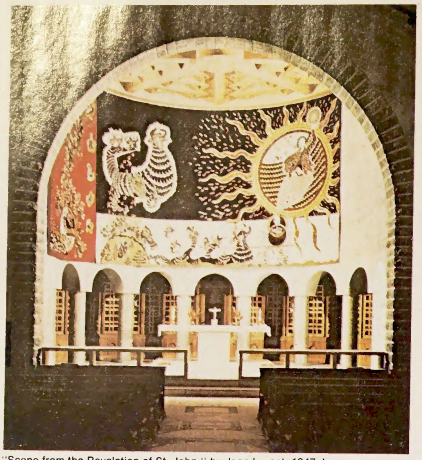




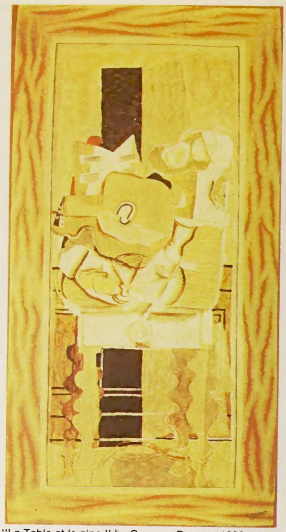
(Above) "The Triumph of Venus," one of four panels of "Marine Triumphs," workshop of Philippe Behagle, late 17th century. In the Banque de France, Paris. 4.35 × 2.60 m.
(Left) "El Cacharrero," woven by the Santa Bárbara Factory after a cartoon by Francisco de Goya, *c.* 1794. In the Palacio Nacional, Madrid. 3.75 × 2.60 m.



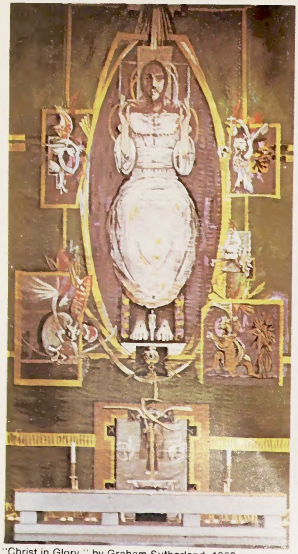
Wall, sofa, and chair tapestries *in situ* woven by the Gobelin workshop after François Boucher's "Loves of the Gods" and "Les Enfants Jardiniers," 1776. In Osterley Park, Middlesex. 3.67 × 6.25 m (wall tapestry).



"Scene from the Revelation of St. John," by Jean Lurcat, 1947. In Notre-Dame de Toute-Grace, Plateau d'Assy, France. 5 × 18 m.



"Le Table et le pipe," by Georges Braque, 1932. In the Arts Club of Chicago. 2.11 × 1.12 m.



"Christ in Glory," by Graham Sutherland, 1962. In Coventry Cathedral, England. 23.94 × 12.05 m.

Tapestries of the 20th century



"Kalota T," woven by the workshop of Tabard after cartoon by Victor Vasarely, 1963. In the Galerie Denise René, Paris. 2.36 × 2.18 m.

Plate 6: By courtesy of (top right) the Arts Club, Chicago, (bottom right) Galerie Denise René, Paris, photographs (top left) Dmitri Kessel, LIFE Magazine © Time Inc., (bottom left) Nicholas Sinnott—Woodmansterne Limited

vidual designers worked on their own looms in the United States and Canada, where most large-scale tapestries continued to be imported from Europe. The Latin-American revival of indigenous folkcrafts aroused interest in tapestry making in Mexico and Panama. South American centres of tapestry art developed in Brazil, Chile, and Colombia.

Modern tapestry design was hindered during the greater part of the 19th century in France by the academic administration of the state factories, although progressive artists began to be affected by the English Arts and Crafts Movement in the late 1880s. The painters Paul Gauguin (1848–1903) and Emile Bernard (1868–1941) were among those who took an interest in tapestry weaving, though they did not actually do tapestry cartoons as did Aristide Maillol (1861–1944). It was not until after World War I that France initiated and led the 20th-century revitalization of tapestry as an art. Many of the great modern artists of the school of Paris—Pablo Picasso (1881–1973), Georges Braque (1882–1962), Henri Matisse (1869–1954), Fernand Léger (1881–1955), Georges Rouault (1871–1958), and Joan Miró (1893–1983), among others—permitted their works to be reproduced in 1932. These reproductions were done with extraordinary fidelity under the supervision of Marie Cuttoli, a Paris connoisseur and promoter of exceptional taste. The Aubusson factory, chosen for this important weaving, became once again a great centre for tapestry. The direct translation of painting into tapestry, however, left little scope for the weaver, and it is the trend begun simultaneously by Jean Lurçat (1892–1966) that may be said to have truly inaugurated the 20th-century tapestry renaissance. Although he began experimenting in 1916, Lurçat's art did not become definitive until the 1930s, when under the influence of Gothic tapestry, particularly the 14th-century "Angers Apocalypse," and in collaboration with François Tabard, master weaver at Aubusson, he formulated the principles that were to make tapestry once again a joint creation between artist and weaver—an art in its own right. No longer merely an imitation painting, tapestry once again exploited the coarser texture and the bolder but more limited range of colours that characterized medieval hangings.

French revitalization of the tapestry art

In 1947 Lurçat founded the important Association des Peintures-Cartonniers de Tapisserie (Association of Cartoon Painters of Tapestry). Also active in this organization were the important French tapestry designers Marc Saint-Saëns and Jean Picart Le Doux, who were Lurçat's foremost disciples. Lurçat was held in great esteem by Dom Robert, a Benedictine monk whose tapestries of poetic fantasy were largely inspired by Persian and medieval European manuscript illumination. Other major French designers of representational compositions were the artists Marcel Gromaire (1892–1971) and Henri Matisse and the architect Le Corbusier (1887–1965).

In the 1950s tapestry designs became increasingly abstract. Among the most notable pieces were those designed by the sculptor and printmaker Henri-Georges Adam (1904–67). Using only black and white, his tapestries are monumental tonal abstractions that reflect his work as an engraver (Figure 106, bottom). The sculptor Jean Arp (1887–1966) and the painter Victor Vasarely are other abstract designers of postwar tapestries.

After World War II the Belgians, influenced by the weaving activity in France during the 1930s, revived their tapestry industry. In 1945 the Forces Murales movement was organized in Tournai by cartoon painters including Louis Deltour, Edmond Dubrunfaut, and Roger Somville, who became the leading designers of Belgian tapestries. This was followed in 1947 by the organization in Tournai of a collective tapestry workshop, the Centre de Rénovation de la Tapisserie, active until 1951. Small workshops continued to flourish in Belgium, especially in the cities of Tournai, Brussels, and Malines.

The renewed international interest in tapestry is clearly related to the austerity of modern architecture. Suitable settings for large-scale wall hangings are provided by the often vast expanses of bare wall surface in contemporary buildings. Le Corbusier not only used tapestries to decorate his architectural interiors but designed them. He frequently referred to tapestries as nomadic murals, recognizing their importance as movable and interchangeable decoration.

In 1962 the first international exhibition of tapestry was held at Lausanne in Switzerland, which after 1965 became an important biennial event. This exhibition clearly demonstrates the tremendous worldwide interest in the medium generated in the middle 20th century as well as indicating the immense variety of tapestry designs, materials, and techniques. (M.J.)

FLORAL DECORATION

Since the earliest days of civilization, man has used floral decorations, composed of living or dried cut-plant materials or artificial facsimiles, to embellish his environment and his person. They have played an important part in folk festivals, religious ceremonials, public celebrations of all kinds, and, of course, courtships. Sophisticated cultures have generally expressed a love for decorating with flowers by carefully arranging them in especially chosen containers, while less sophisticated societies have used them more informally: strewn, made into garlands and wreaths, or casually placed in waterholding vessels without thought of arrangement.

This section covers the elements and principles, the materials, the techniques and forms, and, finally, the historical and stylistic developments of floral design.

Elements and principles of design

The term flower arrangement presupposes the word design. When flowers are placed in containers without thought of design, they remain a bunch of flowers, beautiful in themselves but not making up an arrangement. Line, form, colour, and texture are the basic design elements that are selected, then composed into a harmonious unit based on the principles of design—balance, contrast, rhythm, scale, proportion, harmony, and dominance. Line is provided by branches or slender, steeple-like flowers such as snapdragon, delphinium, and stock. Form and colour are as varied as the plant world itself. Moreover, forms not natural to the plant world can be created for contemporary abstract compositions by bending and manipulating branches, vines, or reeds to enclose space and create new shapes. Texture describes surface quality and can be coarse, as in many-petaled surfaces such as chrysanthemums, or smooth, as in anthuriums, calla lilies, and gladioli. There are many variations between these extremes. Leaves and woody stems also have varied textural qualities.

A flower arrangement includes not only the flowers themselves but the container that holds them and the base on which the container may rest. If an accessory, such as a figurine, is included, that too becomes a part of the total design. The whole composition should relate in textural quality to its frame of reference, which might be a wood or glass table top or a linen cloth, and should be in close harmony with the style of the room for which it was planned, be it Louis XV or Danish modern.

Line and mass

As the components of a design are selected and combined, a silhouette, or arrangement outline, is created. This outline is generally considered most interesting when the spaces in the composition vary in size and shape. Third dimension, or sculptural quality, is accomplished by allowing some of the plant materials in a grouping to extend forward and others to recede. Flower heads turned sideways, or toward the back, for example, break up contour uniformity and draw the eye into and around the composition. When a formal, static quality is sought, the contour is restricted or evenly shaped, often into such graduated forms as a pyramid or mound.

Balance is psychologically important, for an arrangement

that appears to be leaning, top-heavy, or lopsided creates tension in the viewer. (Occasionally, however, as in some modern arrangements, this is the very effect desired.) Colour as well as the actual size of the plant material influences design stability. Dark colour values look heavier than light values; a deep red rose, for example, appears heavier in an arrangement than a pale pink carnation, even though they are the same size. An arrangement in which dark colours are massed at the top and light colours at the bottom can therefore appear top-heavy. Similar flowers placed in identical positions on either side of an imaginary vertical axis create symmetrical balance. If there is an unequal distribution of varying flowers and leaves on either side of the axis but their apparent visual weight is counterbalanced, asymmetrical balance is achieved. This compositional device is more subtle and often more pleasing aesthetically than symmetrical balance, for its effect is less apparently contrived and more varied. Contrasts of light and dark, rough and smooth, large and small, also give variety to the composition. An arrangement generally has a dominant area or centre of visual interest to which the eye returns after examining all aspects of the arrangement. An area of strong colour intensity or very light values, or a rather solid grouping of plant material along the imaginary axis and just above the container's rim, are devices commonly used as compositional centres.

Rhythm, scale, proportion, and harmony

The rhythm of a dynamic, flowing line can be achieved by the graduated repetition of a particular shape, or by the combination of related colour values. Scale indicates relationships: the sizes of plant materials must be suitably related to the size of the container and to each other. Proportion has to do with the organization of amounts and areas; the traditional Japanese rule that an arrangement should be at least one and a half times the height of the container is a generally accepted use of this principle. Proportion also relates to the placement of the arrangement in a setting. A composition is either overpowering or dwarfed if placed on too small or too large a surface or in too small or too large a spatial setting. Harmony is a sense of unity and belonging, one thing with another, that comes with the proper selection of all the components of an arrangement—colour, shape, size, and texture of both plant materials and container.

Materials

Many different kinds of plant materials are used in floral decorations, among them flowers, foliage, grasses, grains, branches, berries, seeds, nuts, cones, fruits, and vegetables. The materials may be living, dried, or artificial. Initially, man was restricted to using native wildings, or uncultivated plants, but as civilization developed over a period of thousands of years, man became less dependent on the seasons and on the resources of the particular region in which he lived. As means of transportation improved and trading grew, plants were introduced from foreign countries and many have since been hybridized to improve or vary shape, size, and colour. In the 20th century the floral decorator has an enormously varied medium in which to create because plant materials can be flown from one part of the world to another. Since the 19th century, when extensive greenhouse cultivation first made it possible to purchase fresh flowers at any time of the year, there have been commercial growers of plant materials who supply the world's floral wholesale markets. The Netherlands, for example, is famous for the 10-mile stretch of greenhouses at Aalsmeer near Amsterdam. In the United States, California and Florida, particularly, have vast areas under cultivation for commercial flowers.

Living plant materials

Dried plant materials are generally used for what is traditionally called a winter bouquet. The cultivated flowers that are often dried are those with a naturally dry, stiff surface quality—such as strawflowers (*Helichrysum bracteatum*), globe amaranth (*Gomphrena*), and statice. North temperate zone wildings picked and preserved for dried arrangements include pearly everlasting, heather, and the sea lavender of salt marshes, as well as goldenrod, orange bittersweet berries, cattails, dock, teasel, and sumac. Many kinds of grasses—pampas, sea oats, millet, and sorghum, for example—are also dried, as are seed-bearing capsules such as the flat paper disks of honesty (*Lunaria*), orange Chinese lanterns (*Physalis*), and the wood roses from the Hawaiian morning glory (*Ipomoea tuberosa*). Other dried materials sometimes used in floral decorations are cones and nuts, long used for making wreaths and festoons for such winter festivals as Christmas; straw, used for Christmas decorations in Sweden and Lithuania; and grains, especially wheat and oats, often arranged in bunches for harvest decorations in Europe and America. Because of their fleshy substance, most fruits and vegetables do not dry well; the main exceptions are gourds, pomegranates, and artichokes.

There are various ways of drying plant materials. Certain garden flowers (among them celosia, blue salvia, globe thistle, alliums, and hydrangeas) can be gathered at their peak of bloom and dried by hanging them upside down in a dark, dry place for several weeks. Flowers may also be individually dried using one of several techniques. A 17th-century Italian writer on horticulture, P. Giovanni Battista Ferrari, described a process of gently burying the flower heads in clean, sun-dried sand and allowing them to remain in a sun-heated place for several months. The same method was used in the 19th century. Later, borax was used, and in the 20th century silica gel, because of its ability to absorb moisture. This solution is gently brushed between and over every petal. Since this method of drying does not preserve the stems, the flower heads must be wired before they are arranged.

Leaves and ferns are dried by pressing. The most delicate pressed flowers and foliage have been composed, mounted, and framed as pictures—a practice especially popular with 19th-century Romantics, who preserved floral souvenirs as sentimental personal memorabilia.

Throughout history and in almost every conceivable medium man has created artificial plant materials. The

By courtesy of the Victoria and Albert Museum, London



Figure 107: Shell-flower arrangement, English, early 19th century. Shells, fastened to the surface of a superstructure, have been used to form an intricate artificial bouquet. In the Victoria and Albert Museum, London.

Artificial plant materials

Chinese fashioned peony blossoms and fruits from semiprecious stones and carved jade leaves, which they assembled into small trees. Gold lotus blossoms were highly treasured in eastern Asia. For European royalty in the late 19th century, the Russian-born jeweller Peter Carl Fabergé (1846–1920) designed exquisite single-stemmed flowers of gold, enamel, gems, and semiprecious stones set in small rock-crystal pots. During the 18th and 19th centuries, the Sèvres porcelain factory in France produced porcelain flowers with stems and leaves of ormolu (a metallic alloy resembling gold). At the same time, the Royal Worcester, Crown Staffordshire, and Royal Doulton factories in England became world-famous for their highly realistic porcelain floral arrangements, which are still made. The Victorians developed a home craft of making and arranging flowers and fruits. Wax, cloth, yarn, feathers, shells, and seeds were used to make the flowers and fruits, which were then either framed or placed under glass domes (Figure 107). Perhaps the most curious of these 19th-century decorations were the wreaths and floral displays made by twisting, knotting, and weaving the hair of one's family and friends around wire supports. Beaded flowers for cemetery and funerary bouquets have been popular in France since the 19th century; and paper flowers for festivals and home decoration have become a major folk art medium in Mexico and Japan. Because of their relatively low cost, durability, and easy maintenance (an occasional washing or dusting), plastic flowers and plants are in such great demand that their production has become an important 20th-century industry. Though still primarily used in public places, plastic plant materials are increasingly found in private homes, especially in the United States.

Techniques

Cut plant materials, especially flowers, need special care and treatment before they are placed in vases. Ideally, flowers are picked some hours before they are arranged and never in the heat of the day. Generally, the bottoms of the stems are cut on a slant, placed in deep tepid water, and kept in a cool place, preferably overnight. Different materials have different conditioning needs. Woody stems are split several inches with pruning shears, then soaked in hot water. Stem ends may be crushed with a mallet instead, but clean cuts make it easier to impale branches on a needle holder. Milky stems, such as those of poppies, poinsettias, and large dahlias, are sealed by placing the tips in boiling water or over a flame for a few seconds. Foliage and flowers are protected from steam and flame by inserting the stems through a hole punched in newspaper, which is then drawn up over them. When arranging flowers, all foliage below the water line must be removed in order to prevent bacterial decay and the resulting unpleasant odour. Since the stems of flowers often seal over while being held in a florist shop or market, they must be recut by the purchaser. Packaged formulas do not aid in revival but are meant to be used during the preliminary soaking period. Roses and woody-stemmed flowers such as chrysanthemums can frequently be revived by recutting and placing them in hot water.

Arranging techniques

Many tall containers can easily display flowers without holding mechanics, but if necessary they can be stuffed with upright pieces of privet or fine evergreens, such as juniper, which are sheared flat across the vase opening. The Japanese *kenzan*, or metal pin holder, usually called a needlepoint holder, is the most generally used mechanical aid. It is held in place with floral clay. In silver vases, melted paraffin is used as a fastener, for, unlike clay, it will not tarnish the container and can be removed easily with hot water. Crumpled chicken wire, or wire netting, is frequently stuffed into vases as an aid to support, and a water-absorbing plastic foam, sold in bricklike blocks, has also become very popular.

Containers and bases

The selection of a suitable container is an individual problem in every arrangement. It is considered a part of the overall design of the arrangement and is related to it in scale, colour, and texture. Its colour must enhance, not compete with, the arrangement. For the same reason many floral decorators prefer to use simply shaped, unadorned vases. The texture of the container is also chosen for compatibility with the floral arrangement. Coarse, heavy plant materials are usually arranged in a substantial container of pottery, pewter, copper, or wood. Delicate flowers and foliage are usually displayed in porcelain, glass, or silver. Fruits and vegetables are often arrayed in wooden or pottery bowls and baskets. The size of the container is also important. If it is too small, the plant materials will overpower it and the arrangement will appear top-heavy. If it is too large, it will not only dwarf the arrangement but will frequently destroy the unity of the composition, dividing the viewer's attention between the floral arrangement and the container. Containers are not used for all arrangements of plant materials. Compositions of driftwood, flowers, fruits, and vegetables are often arranged on a flat base of wood or bamboo, a tray, or slab of wood. To keep them fresh, flowers or foliage used in such an arrangement often are placed in solid-walled pin holders that hold water.

A wooden base frequently completes a composition, since it can add visual weight at the bottom, which assists in balance. The Japanese traditionally use wood or lacquer bases and stands with all arrangements, and a porcelain vase in China was not considered complete without a carved teakwood stand. The stand has both aesthetic and practical advantages: it adds height to a display and prevents moisture stains on furniture or textiles.

Forms of floral decoration

Plant materials are customarily arranged in containers, woven into garlands, and worn or carried for personal adornment. Flower bouquets that are carried include the nosegay and corsage. In the mid-19th century, the nosegay, or posy (a small bunch of mixed flowers), was much in fashion. No well-dressed Victorian lady appeared at a social gathering without carrying one, edged with a paper frill or delicate greens and sometimes inserted into a silver filigree holder (Figure 108). Messages of love were often spelled out in the flowers of the nosegay, for the

Nosegays and corsages

Marc Garanger



Figure 108: Elegant 19th-century use of flowers for personal adornment: "Empress Eugénie," oil on canvas by Édouard Dubufe, 1854. In the Musée National de Versailles et des Trianons, Versailles, France.

"language of flowers" was carefully studied at the time, and courtships progressed through the sending of such floral symbols.

Worn since the 18th century, the corsage has become especially popular in the 20th century. Instead of a nosegay, an admirer frequently sends a lady an orchid, a gardenia, or a small bunch of wired flowers to be worn at the waist, shoulder, or on the wrist, or attached to a handbag and carried. Only the flower heads are used in a corsage. Wires are inserted through the calyx (the usually green or leafy external portion of a flower) and bent to thrust the flowers forward or to the side; then the ends are bound together with tape or ribbon. Leaves or foliage threaded crosswise with wire are usually added. A ribbon bow often completes the corsage.

Sprays are large, flat bouquets of long-stem plant material. They are either carried or placed on caskets or at tombs as commemorative offerings. If the plant material used is short-stemmed, wire is used to add length. The ends of the stems or wire extensions are frequently thrust into a block of moss or stiff plastic foam to secure the arrangement. A blanket of flowers is often laid over a casket at a funeral or over a racehorse in the winner's circle. Blankets are made by stretching burlap over a frame, covering it with a layer of flat fern, and then adding delicate asparagus fern (*Sprengeri*). The fern surface is then covered with flower heads, which are threaded with wire and fastened on the underside of the blanket.

Garlands and wreaths

Garlands are bands of plant materials that have been woven or in some other way attached together; they are not arranged in a container. A circular garland is called a wreath, or if it is worn around the head, a chaplet. Garlands draped in loops are called festoons or swags. The origin of these forms is unknown, but evidence of their use dates from ancient times and is not restricted to any particular culture.

Garlands have been used for many purposes. Ancient Egyptians placed them on mummies. The Greeks used them to decorate their homes, civic places, and temples. For festive occasions the ancient Romans wore garlands of strung rose petals. When these garlands of roses were suspended from ceilings, the conversation that took place beneath them was sub rosa. On European festival days such as Corpus Christi, cattle are bedecked with neck garlands. On Indian holy days, the Hindus take garlands to the temple to be blessed before wearing them; they also hang garlands on the statues of their deities.

In the ancient world it was probably the Romans who most fully developed the ornamental form and use of the festoon. Fine examples are carved in marble on the Ara Pacis or Altar of Peace (13–9 BC) near the Mausoleum of Augustus in Rome. Roman festoons were usually made of fruit, grain, leaves, and flowers. In the late 17th and early 18th centuries it was fashionable, particularly in England, to create artificial festoons over fireplace mantels. Called swags, they were usually carved of wood. Among the most famous are those executed by the English sculptor Grinling Gibbons.

Wreaths have been both worn and displayed. In antiquity the wreath was bestowed upon public officials, athletes, poets, and returning warriors. The ancient Greco-Roman custom of bestowing a laurel crown, or wreath, upon a poet was revived during the Renaissance, especially in Italy. Napoleon chose a laurel wreath of gold for his crown, emulating the emperors of the Roman Empire. At Christmas time since the 19th century, wreaths of evergreens, holly, or pinecones and nuts have been traditionally hung as decorations in northern Europe, the United States, and Canada. In medieval and Tudor England the boar served for Christmas dinner had a wreath of rosemary and bay. During Advent, a period including the four Sundays before Christmas, a wreath with four candles (each symbolizing one of the Advent Sundays) is traditionally hung in Christian homes and churches.

Plant materials have been used for personal adornment in forms other than corsages, nosegays, garlands, and wreaths. Ancient Egyptian wall paintings show women with lotus blossoms in their hair, and today the hibiscus adorns the hair of women of the South Seas. Necklaces of flowers are commonly worn in South and Southeast Asia. In Hawaii, Vanda orchids or velvety frangipani blossoms are strung into long necklaces called leis, the customary gifts of both welcome and farewell.

Many types of dress accessories are decorated with flowers. Staffs ornamented with plant material are seen in ancient art and mentioned in ancient literature. Egyptian servants or standard bearers were often depicted holding staffs of papyrus and lotus blossoms. An attribute of Dionysus and his satyrs was the thyrsus, a staff topped by a pinecone and sometimes further decorated with vine or ivy leaves and grapes. Well-known flowering staffs or rods include those of Aaron, the brother of Moses, and St. Joseph, the earthly father of Jesus.

Flower pictures and floats

Pictorial effects have been achieved by using cut flower heads or petals to create masses of colour that are then worked into patterns. The traditional carpet of flowers laid down on the Via Livia in Genzano, Italy, for the feast of Corpus Christi is incredibly intricate and colourful. Figures of angels, madonnas, and saints, geometric designs, and coats of arms are worked out with flower petals to form a carpet over which the religious procession passes. Mexicans frequently carpet their churches with mosaics of wild flowers, and in The Netherlands during tulip time flower pictures are made for competition. About 12 feet square, made for the most part of tulip and hyacinth blossoms, they are designed on inclined backgrounds for better visibility. Some of these pictures are three-dimensional.

For centuries flower-covered floats have been used in parades. The Italian painter and architect Giorgio Vasari (1511–74) described 21 garland-decorated floats he designed for a pageant in Florence. The most famous of modern floral parades is the Tournament of Roses parade held on New Year's Day at Pasadena, California. Floats up to 50 feet (15 metres) in length are constructed over the chassis of motor vehicles. Rough framework is covered with chicken wire shaped and sprayed with a polyvinyl coating. Flower heads are attached with either glue or wire.

SCALA—Art Resource



Figure 109: Figures wearing wreaths of leaves and flowers, "Heracles and Telephos Before the Personification of Arcadia," Roman wall painting from Herculaneum (c. 1st century AD), after a Hellenistic original, early 2nd century BC. In the Museo Archeologico Nazionale, Naples.

Historical and stylistic developments

WESTERN

Ancient world. There is evidence through painting and sculpture that during the Old Kingdom (*c.* 2686–*c.* 2160



Figure 110: (Top) Stylized bouquets of lotus and buds bound with rows of petals and berries, "Apuy and His Wife Receiving Offerings," tempera copy of an Egyptian wall painting from the tomb of Apuy at Thebes. In the Metropolitan Museum of Art, New York. (Bottom) Earliest representation of mixed flowers artfully arranged in a container, "Basket of Flowers," Roman mosaic, 2nd century. The basket holds Roman hyacinths, roses, tulips, red carnations, a double anemone, and a blue morning glory. In the Vatican Museum, Rome.
By courtesy of (top) the Metropolitan Museum of Art, New York photograph, (bottom) SCALA—Art Resource

BC) the Egyptians placed flowers in vases. In the tomb of Perneb bas-relief carvings show lotus blossoms and buds alternately arranged in flared bowls that were set upon banquet tables or carried in processions. Paintings of functional vases with spouts designed to support the heavy-headed lotus flower are found in the tombs of Beni Hasan (*c.* 2500 BC). Formal bouquets of lotus and berries offered to the dead are represented in the paintings from the tomb of Apuy at Thebes (Figure 110, top). Garlands and wreaths, floral headdresses, and collars were woven. Because of the formalized rules of Egyptian art, the lotus (*Nymphaea*), sacred to the goddess Isis, and papyrus, both of which were easily conventionalized, were the plant materials depicted almost exclusively for 2,000 years. During the Ptolemaic era (305–30 BC) perfume recipes, flower garlands found on mummies, and Greek and Roman writings reveal a more varied native plant life and show that foreign plants had been introduced, most notably the rose.

The ancient Greeks' love of flowers was expressed mainly in the making and wearing of wreaths and garlands (Figure 109). Vase paintings, temple friezes, and architectural ornamentation all illustrate their widespread use. They were also frequently mentioned in Greek literature. The techniques of making garlands and wreaths, the most appropriate plant materials, and the proper time and way to wear or display them, were the subjects of several treatises. Fruits and vegetables mounded in baskets or spilling in profusion out of a cornucopia were types of arrangements used for religious offerings.

The earliest depiction of mixed cut flowers, artfully arranged in a container, is a mosaic dating from the early 2nd century AD of a basket of flowers from the emperor Hadrian's villa at Tivoli near Rome (Figure 110, bottom). Garlands and wreaths continued to be popular among the Romans, as did displays of fruits and vegetables in cornucopias and baskets.

Middle Ages. Little evidence remains of floral decoration in early medieval Europe. In the mosaics of Ravenna, the Byzantines depicted highly contrived formal compositions. Symmetrical, with an emphasis on height, these arrangements were usually spires of foliage with regularly placed clusters of flowers or fruit.

Medieval plant symbolism

Illuminated manuscripts of the Gothic period (from the 13th century to the 15th) occasionally include simple floral bouquets holding symbolic flowers. This was a time of intense religious fervour, and plant symbolism assumed great importance. There was both a liturgical and a secular language of flowers. In the church, for example, the rose symbolized the Virgin; in the chivalric courts, passionate love. Usually plant materials were casually placed in utilitarian containers such as earthenware jugs, bottles, glass tumblers, and in majolica, or glazed and enamelled pottery, drug jars called albarelli. The still life in the foreground of the open centre panel of the "Portinari Altarpiece" by the Flemish painter Hugo van der Goes (Figure 111) is an illustration of this type of arrangement. Metal ewers often held Madonna lilies (*Lilium candidum*), as in the 15th-century painting "The Annunciation" by Rogier van der Weyden (Metropolitan Museum of Art, New York).

15th and 16th centuries. Floral decorations became more studied and elaborate during the Renaissance period of the 15th and 16th centuries. The revival of interest in antiquity influenced the widespread use of garlands and wreaths in Renaissance Europe, especially in Italy. They were popular for pageants and feasts as well as for decorating houses and churches, and were commonly depicted in the art of the time. Among the most notable examples are the terra-cotta wreaths that framed the decorative ceramic plaques and reliefs made by the della Robbia family in the

SCALA—Art Resource



Figure 111: Symbolic use of flowers, "Portinari Altarpiece" (detail from the central panel) by Hugo van der Goes, *c.* 1476. The scattered violets indicate Christ's humility; the columbine flowers represent the seven gifts of the Holy Spirit with which Christ was endowed at birth. The fleur-de-lis indicates royalty, and the flowers in the albarello are in royal colours, for Christ was of the royal line of David. In the Uffizi, Florence.



Figure 112: Formal and elaborate arrangement of fruits, vegetables, and flowers characteristic of Renaissance floral decoration, "Madonna of Victory," altarpiece by Andrea Mantegna, 1495. In the Louvre, Paris.
Giraudon—Art Resource

late 15th century, and the garlands of flowers, fruits, and vegetables in the paintings of such northern Italian masters as Andrea Mantegna and Carlo Crivelli (Figure 112). Cut-plant materials were generally arranged in either high sparse bouquets or tight low bunches. There were also pyramidal compositions in pedestal vases, such as those in the background of the painting "Virgin and Child and St. John" (Borghese Gallery, Rome) by the Florentine artist Sandro Botticelli. Arrangements of fruits and vegetables on salvers or in baskets also were popular.

17th century. The arrangement of plant materials truly became an art and an important decorative device in the 17th century. During this period of worldwide exploration, colonization, and commerce, new plants were introduced into Europe, where an avid interest in horticulture developed. Still-life paintings of the late 16th, 17th, and early 18th centuries reveal what a great variety of plants there was in the gardens of Europe. Beginning with Jan Brueghel (called "Velvet Brueghel"; 1568–1625), a tradition of flower painting developed in Flanders and Holland, which culminated with the works of Jan van Huysum (1682–1749). The canvases of the many hundreds of still-life painters of the period are valuable source material for the student of the history of floral decorations and gardens. They must, however, be considered as idealized compositions and not as literal translations onto canvas of actual bouquets. Early 17th-century pictures, particularly those of Jan Brueghel, who painted one-of-a-kind arrangements, seemed most interested in displaying the content of the garden itself. Depictions of later 17th-century bouquets show profuse arrangements that reflect the sensuality and exuberance of the Baroque style. Curvilinear elements such as sinuous S curves are other Baroque devices of design used to createagrandiloquent, dramatic compositions. The massed bouquets of the Baroque period are studies in dominance, contrast, rhythm, and sculptural effect. The eye is drawn around and into the bouquets by the turning of flower heads, the reversing of leaves, and the curving of graceful flower stems.

National styles of arrangement

The French style of the Louis XIV period (1643–1715) is best exemplified in the flower engravings of Jean-Baptiste Monnoyer. The plates for his famous portfolio *Le Livre de toutes sortes de fleurs d'après nature (Book of All Kinds of Flowers from Nature)* accurately portray flowers from a horticultural standpoint and at the same time show prototypes of display. These floral arrangements are freer and more airy than those of the Low Countries and yet suggest Baroque opulence. *Flora ouero cultura di fiori* ("Flora: The Cultivation of Flowers"), a renowned garden book published in Rome in 1633 by the horticulturist P. Giovanni Battista Ferrari, illustrates the styles of floral displays preferred by the Italians and also describes arranging techniques and devices. Among the ingenious devices illustrated is a vase with holes in its removable top that made it easy to arrange flowers and change water.

18th century. The floral arrangements of the early 18th century were dominated by French and English taste. In France, cultural and social life centred in the intimate rooms of Parisian town houses rather than in the vast rooms and halls of Louis XIV's Versailles palace. Bouquets, therefore, were comparatively small, to be in scale with their setting. The more delicate colouring and lighter visual weight of these arrangements can be attributed in part to feminine taste, which decidedly influenced the Rococo style (Figure 113). Personal and charming, the Rococo bouquet and its variations remained popular into the 20th century. English bouquets of the corresponding Georgian period were often more profuse than the Rococo. Many books written to catalog the wide variety of plant materials available in 18th-century England gave incidental information on how to care for and display them. One of the best known of these works is the two-volume *Gardeners Dictionary* by the horticulturist Philip Miller. In it he mentions dried bouquets and chimney flowers. It was customary in English homes to arrange flowers and branches in the hearth during the summer months when the fireplace was not in use. These arrangements were referred to as "bough pots." The best known English illustrations of Georgian flower arrangements are those designed by the Flemish artist Peter Casteels for a nursery catalog called *The Twelve Months of Flowers* (1730). Since the flowers in each bouquet are numbered and keyed to a list at the bottom of the plate, and are one-of-a-kind collections, they are not truly representative of live arrangements. Jacob van Huysum's monthly paintings display flowers more naturally. Both series are invaluable as source material for garden flowers.

The Neoclassical period of the late 18th and early 19th centuries brought about a revival of wreaths and garlands in the style of Greco-Roman antiquity. Floral bouquets were arranged in vases of classical severity.

19th century. The interest of the 19th-century Romantics in nature made floral arrangements an important part of a decorative scheme. With the advent of the clipper ship more exotic plant materials were introduced into Europe and the United States. From China came new varieties of chrysanthemums, bleeding heart, rhododendrons, and azaleas; from South Africa, the gladiolus, freesia, and pelargoniums; and from Mexico, the dahlia, gloxinia, and fuchsia. Many old garden favourites were greatly improved as a result of widespread scientific interest in horticulture and botany. The Industrial Revolution made it possible to manufacture a great variety of economically priced ceramic and glass containers. Artificial flowers were extremely popular and were made in many different materials in both home and factory.

Victorian arrangements

The books and magazines of the Victorian age agreed that the art of arranging flowers was an accomplishment all young ladies should acquire. Except for the single flower in the small bud vase, the most popular style of Victorian arrangement was a tightly compact mass of flowers,



Figure 113: (Left) An exuberant and dramatic Baroque flower arrangement, "Flowers in a Vase," oil on panel by Jan van Huysum, 1726. In the Wallace Collection, London. (Right) An intimate and delicate bouquet of the 18th century, "A Vase of Flowers," oil on canvas by J.-B.-S. Chardin (1699–1779). In the National Gallery of Scotland, Edinburgh.

(Left) by permisson of the trustees of the Wallace Collection, London, (right) by courtesy of the National Gallery of Scotland, Edinburgh

greens, grasses, and ferns. The two-level epergne, with a flared top for flowers and lower tier for fruit, frequently was used for the centre of the dining table. Since the flowers selected were usually of a brilliant hue, strong colour contrast was a characteristic of Victorian arrangements. These gay floral groupings, however, were usually softened by ferns and other kinds of foliage.

20th century. The book *Flower Decoration in the House* (1907) greatly influenced the development of 20th-century floral decoration as an art. The author was Gertrude Jekyll, already notable in the gardening world. For a long time, floral decoration in big houses had been the charge of the head gardeners or the local florists; in smaller houses, the charge of the mistress of the house. In any case, arrangement was done with varying degrees of skill and little guidance. With Gertrude Jekyll's book, the idea that flower decorations actually could be planned and designed in such a way as to heighten the quality of a room came to be widely accepted. Interior decorators added their specialized knowledge to the practical expression of this view.

The rise of the women's Garden Club movement in the 1930s and the growth of flower shows led to establishing definite rules for arrangement, especially in the United States. The classic Japanese rules of design (see below) were adopted, and others were formulated. Three main types of arrangement were recognized—the mass, the line, and the combination line-mass. Emphasis was placed on design shapes such as the crescent, or Hogarth curve, and colour studies in related or contrasting harmonies. In exhibitions thematic compositions were popular, and often arrangements interpreted abstract ideas, emotions, places, and natural phenomena. Naturalistic compositions with just a few flowers made use of stones, moss, and branches or driftwood with striking line interest (Figure 114). In the mid-20th century flower arranging tended to follow contemporary art trends. A Japanese revolt against traditional aesthetic canons also had great influence on Western development of free-style arrangements that reject naturalism and are often unconventional in their placement and use of treated material. Traditional principles of visual design are often rejected in such modern arrangements.

Assemblages and abstract arrangements

Assemblages of such diverse elements as scrap metal, rope, and plastic are composed with a minimum of plant material. Transition and rhythm yield to heightened contrast. Space is important, and new forms are created by bending plant material to create new shapes. Psychological tension is created by upsetting balance and symmetry.

EASTERN

China and Korea. The ancient Chinese could enjoy and feel themselves at one with the growth, maturity, and decline of a few flowers or a branch. The floral expressions of the Chinese have traditionally been based on the Confucian art of contemplation, the Buddhist principle of preservation, and Taoist symbolism (Figure 115). For the Confucian, a floral arrangement was philosophically contemplated both as a symbol of organic existence and for its aesthetic aspects. Buddhists used flowers sparingly because of their religious doctrine prohibiting the taking of life. At least since the T'ang dynasty (AD 618–907), flowers have been placed on temple altars in a *ku*, an ancient bronze

Constance Spry Ltd



Figure 114: Naturalistic 20th-century composition of spring flowers: daffodils, daisies, primroses, and pasqueflowers arranged with moss and heath in a frame of bracket fungus.



Figure 115: "Western Queen Mother" (detail of flower border), silk and metal thread embroidered on silk, Chinese, 19th century. The plant material includes (left to right) plum, pine, citron, lotus, orchid, magnolia, peony, peach, chrysanthemum, *Osmanthus*, and *Narcissus tazetta*. In the Metropolitan Museum of Art, New York.

By courtesy of the Metropolitan Museum of Art, New York, bequest of William Crawford, 1929

ceremonial wine beaker dating from the Shang dynasty (18th to 12th century BC) whose shape was translated into porcelain in later dynasties. Hua Hsien, the flower goddesses of the Taoists, have traditionally been represented carrying flower-filled baskets. In Taoist symbolism, the four seasons were denoted by the white plum blossom of winter, the peony of spring, the lotus of summer, and the chrysanthemum of autumn. Each month also had its own flower. Longevity in plant arrangements was symbolized by pine, bamboo, and the long-lasting *ling chih* fungus. New Year floral displays featured the paper-white narcissus, and the tree peony (*Paeonia moutan*), designated the "king of flowers," was used to symbolize good fortune.

Usually the floral arrangements of the Chinese, like those of the Koreans, appear less obviously contrived than those of the Japanese. A composition frequently will be made of two or more arrangements in containers of different heights and shapes, often grouped with rocks or decorative objects. Chinese bouquets in baskets have a quality reminiscent of Western floral arrangements.

Japan. The arrangement of flowers in Japan is an elaborate and unique art, with highly developed conventions and complex symbolism. The art developed from the custom of offering flowers to the Buddha and was introduced into Japan early in the 7th century by Ono No Imoko, Japanese ambassador to China, who founded the first and oldest school of floral art, the Ikenobō. All the later masters of the Ikenobō school are his descendants. Most important among the earliest styles was the *mitsu-gusoku*, an arrangement of three or five articles often consisting of an incense burner, a candlestick in the form of a stork, and a vase of flowers. These were usually displayed before pictures of the Buddha or of founders of Buddhist sects.

Early Japanese styles

Early styles were known as *tatebana*, standing flowers; from these developed a more massive and elaborate style, *rikka* (which also means standing flowers), introduced by the Ikenobō master Senkei around 1460 (Figure 116). The early *rikka* style symbolized the mythical Mt. Meru of Buddhist cosmology. *Rikka* represented seven elements: peak, waterfall, hill, foot of the mountain, and the town, and the division of the whole into *in* (shade) and *yō* (sun). (In Chinese the characters for *in* and *yō* are read *yin* and *yang*, the passive or female and the active or male principles.) Formal *rikka* is arranged out of nine main branches and some accessory ones. Three branches are placed so that their tips form a triangle with unequal sides. From this pattern all later styles of Japanese floral art developed.

In the early 18th century a three-branch, asymmetrical style, *shōka*, evolved from the *rikka* and was cultivated by the Ikenobō school. *Shōka* is written with Japanese characters meaning living flowers. These characters can also be read *seika* and *ikebana*; *seika* is the preferred reading by some schools, while *ikebana* today is the general term applied to any style of Japanese floral art. Up to the advent of *shōka* all styles of arrangements other than *rikka* had been known as *nageire*, meaning to throw, or fling into. This term was confined to arrangements in tall vases, and *heika*, vase flowers, is preferred to *nageire* by some schools. *Shōka* utilized three main branches, and emulated the natural growth of plant life. This illusion of growth was achieved by using buds, foliage, and blossoms; by superimposing stems as they emerged from the container; by turning up the tip ends of branches unless of a naturally drooping kind; and by placing tree branches above flowers and mountain material above that of the lowland. All combinations were seasonally correct. Uneven numbers of materials were always used, and rules of proportion dictated that plant material be at least one and one half times the height of the container. By the late 18th and early 19th centuries the *shōka* style had supplanted the *rikka* in popularity and many new schools flourished, including Enshūryū, Koryū, Kōdōryu, and Mishōryū. All these new schools utilized the three-branch form but adopted different nomenclatures for them.

By courtesy of the Tokyo National Museum



Figure 116: Traditional *rikka* flower arrangement of nine different plant materials. In the Tokyo National Museum.

Western flowers were introduced into Japan following the Meiji Restoration (1868). The flower master Ohārā Unshin, who established the Ohara school (early 20th century), devised for them a new container, based on the low bowls used for dwarfed plants. This new style, known as *moribana* (heaped-up flowers), permitted greater freedom in the choice and placement of materials. A variation was the creation of small realistic landscapes called *shakei*, sometimes referred to as memory sketches. In these, exposed water surface was a part of the design. In 1930 a group of art critics and flower masters proclaimed a new style of floral art called *zen'ei ikebana* (avant-garde flowers), free of all ties with the past. Foremost in this group was the Ikenobō master Teshigahara Sōfū (1900–79), who had founded the Sōgetsu school in 1927. The

new style emphasized free expression. It utilized all forms of plant life, living and dead, and elements that had been previously avoided, such as bits of iron, brass, vinyl, stone, scrap metal, plastic, and feathers. Vines and branches were bleached and painted and even used upside down. Stems were crossed, even numbers of materials were used, and containers were often crude and exotic in shape.

Until 1868 Japanese flower arrangement was generally a man's avocation, engaged in primarily by Buddhist priests, warriors, and the nobility. Following the Meiji restoration and particularly after the beginning of the 20th century, it was taken up by large numbers of women. Men, however, still head most of the principal schools.

The total number of schools that teach floral decoration throughout Japan in the 20th century is believed to number from 2,000 to 3,000, varying in size from several thousand to millions of adherents. Each school has its own rules of arrangement, though styles may differ only slightly from one another. All arrangements are asymmetrical and achieve a three-dimensional effect. The traditional styles are still taught, many with modern variations, but the bolder, less restrained, and unconventional free-style forms of arrangement now seem to be the most popular. The material used in Japanese floral arrangements is held in position by various artifices, the most popular of which are the *kubari,* forked twig, and the *kenzan,* needlepoint holder.

Japanese flower arranging has influenced that of the West considerably, particularly in the mid-20th century. Many popularizations of the art have flourished in the United States.

OTHER CULTURES

Outside the West and the Far East, the arranging of plant materials was more a casual part of everyday life than a formally recognized medium of artistic expression. The elaborate stylistic traditions evolved and formulated in the West and Far East through centuries of sophisticated creative activity are rarely found, therefore, in other cultures. In the Islāmic world, for example, simple, modestly scaled arrangements predominated: sparse, symmetrically arranged bouquets; casually grouped bunches of flowers; or blossoms floating on liquid surfaces. The garlands made in India for adorning home, temple, statuary, and man himself were simpler than the bouquet or arranged floral materials found in the more aesthetically complex traditions of the West and Far East. Also in contrast to these artistically self-conscious arrangements are the stiff, mounded groupings of plant materials made for festivals in Southeast Asia. (Ju.S.B.)

POTTERY

Pottery in its widest sense includes all objects made from clay and hardened by fire: earthenware, stoneware, and porcelain. This article deals with pottery as art and craft, covering the kinds of pottery, the processes and techniques of forming and decorating it, and, finally, the history of pottery. Since the subject is so vast, the article will deal only with museum, or art, pottery. Artifacts used for purely utilitarian purposes are discussed in articles such as AFRICAN ARTS.

Kinds, processes, and techniques

Clay and firing

Clay, the basic material of pottery, has two distinctive characteristics: it is plastic (*i.e.*, it can be molded and will retain the shape imposed upon it); and it hardens on firing to form a brittle but otherwise virtually indestructible material that is not attacked by any of the agents that corrode metals or organic materials. Firing also protects the clay body against the effects of water. If a sun-dried clay vessel is filled with water, it will eventually collapse, but, if it is heated, chemical changes that begin to take place at about 900° F (500° C) preclude a return to the plastic state no matter how much water is later in contact with it. Clay is a refractory substance; it will vitrify only at temperatures of about 2,900° F (1,600° C). If it is mixed with a substance that will vitrify at a lower temperature (about 2,200° F, or 1,200° C) and the mixture is subjected to heat of this order, the clay will hold the object in shape while the other substance vitrifies. This forms a nonporous, opaque body known as stoneware. When feldspar or soapstone (steatite) is added to the clay and exposed to a temperature of 2,000° to 2,650° F (1,100° to 1,450° C), the product becomes translucent and is known as porcelain. In this section, earthenware is used to denote all pottery substances that are not vitrified and are therefore slightly porous and coarser than vitrified materials.

The line of demarcation between the two classes of vitrified materials—stoneware and porcelain—is extremely vague. In the Western world, porcelain is usually defined as a translucent substance—when held to the light most porcelain does have this property—and stoneware is regarded as partially vitrified material that is not translucent. The Chinese, on the other hand, define porcelain as any ceramic material that will give a ringing tone when tapped. None of these definitions is completely satisfactory: for instance, some thinly potted stonewares are slightly translucent if they have been fired at a high temperature, whereas some heavily potted porcelains are opaque. Therefore, the application of the terms is often a matter of personal preference and should be regarded as descriptive, not definitive.

KINDS OF POTTERY

Earthenware. Earthenware was the first kind of pottery made, dating back about 9,000 years. In the 20th century, it is still widely used.

The earthenware body varies in colour from buff to dark red and from gray to black (see below *Firing*). The body can be covered or decorated with slip (a mixture of clay and water in a creamlike consistency, used for adhesive and casting as well as for decoration), with a clear glaze, or with an opaque tin glaze. Tin-glazed earthenware is usually called maiolica, faience, or delft (see below *Decorative glazing*). If the clear-glazed earthenware body is a cream colour, it is called creamware. Much of the commercial earthenware produced in the second half of the 20th century is heat- and cold-proof and can thus be used for cooking and freezing as well as for serving.

Stoneware. Stoneware is very hard and, although sometimes translucent, usually opaque. The colour of the body varies considerably; it can be red, brown, gray, white, or black.

Fine white stoneware was made in China as early as 1400 BC (Shang dynasty). In Korea, stoneware was first made during the Silla dynasty (57 BC–AD 935); in Japan, during the 13th century (Kamakura period). The first production of stoneware in Europe was in 16th-century Germany. When tea was first imported to Europe from China in the 17th century, each chest was accompanied by a red stoneware pot made at the I-hsing kilns in Kiang-su province. This ware was copied in Germany, the Netherlands, and England. At the end of the 17th century, English potters made a salt-glazed white stoneware that was regarded by them as a substitute for porcelain (see below *Decorative glazing*). In the 18th century, the Englishman Josiah Wedgwood made a black stoneware called basaltes and a white stoneware (coloured with metallic oxides) called jasper. A fine white stoneware, called Ironstone china, was introduced in England early in the 19th century. In the 20th century, stoneware is used mostly by artist-potters, such as Bernard Leach and his followers.

The first porcelain

Porcelain. Porcelain was first made in a primitive form in China during the T'ang dynasty (AD 618–907). The kind most familiar in the West was not manufactured until the Yüan dynasty (AD 1279–1368). It was made from kaolin (white china clay) and petuntse (a feldspathic rock), the latter being ground to powder and mixed with the clay. During the firing, which took place at a temperature of

about 2,650° F (1,450° C), the petuntse vitrified, while the refractory clay ensured that the vessel retained its shape.

In medieval times isolated specimens of Chinese porcelain found their way to Europe, where they were much prized, principally because of their translucency. European potters made numerous attempts to imitate them, and, since at that time there was no exact body of chemical and physical knowledge whereby the porcelain could be analyzed and then synthesized, experiments proceeded strictly by analogy. The only manufactured translucent substance then known was glass, and it was perhaps inevitable that glass made opaque with tin oxide (the German *Milchglas,* or milk glass, for example) should have been used as a substitute for porcelain. The nature of glass, however, made it impossible to shape it by any of the means used by the potter, and a mixture of clay and ground glass was eventually tried. Porcelain made in this way resembles that of the Chinese only superficially and is always termed soft, or artificial, porcelain. The date and place of the first attempt to make soft porcelain are debatable, but some Middle Eastern pottery of the 12th century was made from glaze material mixed with clay and is occasionally translucent (see below *Islāmic pottery: Egypt*). Much the same formula was employed with a measure of success in Florence about 1575 at workshops under the patronage of Duke Francesco de'Medici. No further attempts of any kind appear to have been made until the mid-17th century, when Claude and François Révérend, Paris importers of Dutch pottery, were granted a monopoly of porcelain manufacture in France. It is not known whether they succeeded in making it or not, but, certainly by the end of the 17th century, porcelain was being made in quantity, this time by a factory at Saint-Cloud, near Paris.

The secret of true, or hard, porcelain similar to that of China was not discovered until about 1707 in Saxony, when Ehrenfried Walter von Tschirnhaus, assisted by an alchemist called Johann Friedrich Böttger, substituted ground feldspathic rock for the ground glass in the soft porcelain formula. Soft porcelain, always regarded as a substitute for hard porcelain, was progressively discontinued because it was uneconomic; kiln wastage was excessive, occasionally rising to nine-tenths of the total.

Hard and soft porcelain

The terms soft and hard porcelain refer to the soft firing (about 2,200° F, or 1,200° C) necessary for the first, and the hard firing (about 2,650° F, or 1,450° C) necessary for the second. By coincidence they apply also to the physical properties of the two substances: for example, soft porcelain can be cut with a file, whereas hard porcelain cannot. This is sometimes used as a test for the nature of the body.

In the course of experiments in England during the 18th century, a type of soft porcelain was made in which bone ash (a calcium phosphate made by roasting the bones of cattle and grinding them to a fine powder) was added to the ground glass. Josiah Spode the Second later added this bone ash to the true, hard porcelain formula, and the resulting body, known as bone china, has since become the standard English porcelain. Hard porcelain is strong, but its vitreous nature causes it to chip fairly easily and, unless especially treated, it is usually tinged slightly with blue or gray. Bone china is slightly easier to manufacture. It is strong, does not chip easily, and the bone ash confers an ivory-white appearance widely regarded as desirable. Generally, bone china is most popular for table services in England and the United States, while hard porcelain is preferred on the European continent.

FORMING PROCESSES AND TECHNIQUES

Raw clay consists primarily of true clay particles and undecomposed feldspar mixed with other components of the igneous rocks from which it was derived, usually appreciable quantities of quartz and small quantities of mica, iron oxides, and other substances. The composition and thus the behaviour and plasticity of clays from different sources are therefore slightly different. Except for coarse earthenwares, which can be made from clay as it is found in the earth, pottery is made from special clays plus other materials mixed to achieve the desired results. The mixture is called the clay body, or batch.

To prepare the batch, the ingredients are combined with water and reduced to the desired degree of fineness. The surplus water is then removed.

Shaping the clay. The earliest vessels were modelled by hand, using the finger and thumb, a method employed still by the Japanese to make *raku* teabowls. Flat slabs of clay luted together (using clay slip as an adhesive) were employed to make square or oblong vessels, and the slabs could be formed into a cylinder and provided with a flat base by the same means. Coiled pottery was an early development. Long rolls of clay were coiled in a circle, layer upon layer, until the approximate shape had been attained; the walls of the vessel were then finished by scraping and smoothing. Some remarkably fine early pots were made in this way.

Introduction of the potter's wheel

It is impossible to say when the potter's wheel, which is a difficult tool and needs long apprenticeship, was introduced. A pot cannot be made by hand modelling or coiling without the potter either turning it or moving around it, and, as turning involves the least expenditure of human effort, it would obviously be preferred. The development of the slow, or hand-turned, wheel as an adjunct to pottery manufacture led eventually to the introduction of the kick wheel, rotated by foot, which became the potter's principal tool. The potter throws the clay onto a rapidly rotating disc and shapes his pot by manipulating it with both hands. This is a considerable feat of manual dexterity that leads to much greater exactness and symmetry of form. Perhaps the most skillful of all potters have been the Chinese. Excellent examples of their virtuosity are the double-gourd vases, made from the 16th century onward, which were turned in separate sections and afterward joined together. By the 18th century the wheel was no longer turned by the potter's foot but by small boys, and since the 19th century the motive power has been mechanical.

Jollying, or jiggering, is the mechanical adaptation of wheel throwing and is used where mass production or duplication of the same shape—particularly cups and plates—is required. The jolly, or jigger, was introduced during the 18th century. It is similar to the wheel in appearance except that the head consists of a plaster mold shaped like the inside of an object, such as a plate. As it revolves, the interior of the plate is shaped by pressing the clay against the head, while the exterior, including the footring, is shaped by a profile (a flat piece of metal cut to the contour of the underside of the plate) brought into contact with the clay. Machines that make both cups and plates automatically on this principle were introduced in the 20th century. Small parts, such as cup handles, are made separately by pressing clay into molds and are subsequently attached to the vessel by luting.

Molding

One of the earliest methods of shaping clay was molding. Pots were made by smearing clay around the inside of a basket or coarsely woven sack. The matrix was consumed during firing, leaving the finished pot with the impression of the weave on the exterior. A more advanced method, used by the Greeks and others, is to press the pottery body into molds of fired clay. Though the early molds were comparatively simple, they later became more complex, a tendency best seen in those molds used for the manufacture of pottery figures. The unglazed earthenware figures of Tanagra (Boeotia, central Greece) were first modelled by hand, then molds of whole figures were used, and finally the components—arms, legs, heads, and torsos—were all molded separately. The parts were often regarded as interchangeable, so that a variety of models could be constructed from a limited number of components. No improvement on this method of manufacture had been devised by the 20th century: the European porcelain factories make their figures in precisely the same way.

Plaster of paris molds were introduced into Staffordshire about 1745. They enabled vessels to be cast in slip, for when the slip was poured into the mold the plaster absorbed the water from it, thus leaving a layer of clay on the surface of the mold. When this layer had reached a sufficient strength and thickness, the surplus slip was poured off, the cast removed and fired, and the mold used again. This method is still in common use.

Drying, turning, and firing. Newly shaped articles were

formerly allowed to dry slowly in the atmosphere. In 20th century pottery factories, this stage is speeded up by the introduction of automatic dryers, often in the form of hot, dry tunnels through which the ware passes on a conveyor belt.

Turning is the process of finishing the greenware (unfired ware) after it has dried to leather hardness. The technique is used to smooth and finish footrings on wheel-thrown wares or undercut places on molded or jiggered pieces. It is usually done on the potter's wheel or jigger as the ware revolves. Lathe turning, like most hand operations, was tending to disappear in the mid-20th century except on the more ornamental and expensive objects.

The firing process

The earliest vessels, which were sun-dried but not fired, could be used only for storing cereals and similar dry materials. If a sun-dried clay vessel is filled with water it absorbs the liquid, becomes very soft, and eventually collapses; but if it is heated, chemical changes that begin to take place at about 900° F (500° C) preclude a return to the plastic state.

After thorough drying, the pottery is fired in a kiln. In primitive pottery making, the objects were simply stacked in a shallow depression or hole in the ground, and a pyre of wood was built over them. Later, coal- or wood-fired ovens became almost universal. In the 20th century both gas and electricity are used as fuels. Many improvements have been made in the design of intermittent kilns, in which the ware is stacked when cold and then raised to the desired temperature. These kilns are extravagant of fuel, however, and are awkward to fill or empty if they have not had time to cool completely. For these reasons they are being replaced by continuous kilns, the most economical and successful of which is the tunnel kiln. The wares are conveyed slowly from a comparatively cool region at the entrance to the full heat in the centre. As they near the exit after firing, they cool gradually.

The atmosphere in the kiln at the time of firing, as well as the composition of the clay body, determines the colour of the fired earthenware pot. Iron is ubiquitous in earthenware clay, and under the usual firing conditions it oxidizes, giving a colour ranging from buff to dark red according to the amount present. In a reducing atmosphere (*i.e.*, one where a limited supply of air causes the presence of carbon monoxide) the iron gives a colour varying from gray to black, although a dark colour may also occur as a result of the action of smoke. Both of the colours that result from iron in the clay can be seen in the black-topped vases of predynastic Egypt.

DECORATING PROCESSES AND TECHNIQUES

Impressing and stamping. Even the earliest pottery was usually embellished in one way or another. One of the earliest methods of decoration was impressing the raw clay. Finger marks were sometimes used, as well as impressions from rope (Japanese Jōmon ware of the 1st millennium BC) or from a beater bound with straw (used to shape the pot in conjunction with a pad held inside it). Basketwork patterns are found on pots molded over baskets and are sometimes imitated on pots made by other methods.

Relief decoration

The addition of separately modelled decoration, known as applied ornament (or appliqué), such as knops (ornamental knobs) or the reliefs on Wedgwood jasperware, came somewhat later. The earliest known examples are found on Mediterranean pottery made at the beginning of the 1st millennium. Raised designs are also produced by pressing out the wall of the vessel from inside, as in the Roman pottery known as terra sigillata, a technique that resembles the repoussé method adopted by metalworkers. Relief ornament was also executed—by the Etruscans, for example—by rolling a cylinder with the design recessed in intaglio over the soft clay, the principle being the same as that used to make Babylonian cylinder seals.

Incising, sgraffito, carving, and piercing. The most primitive kind of decoration is that which is incised into the raw clay with a pointed stick or with the thumbnail, chevrons (inverted v's) being a particularly common motif. Incised designs on a dark body are sometimes filled with lime, which effectively accents the decoration. Examples can be seen in some early work from Cyprus and in some comparatively modern work from primitive tribes. Decoration that has been engraved after firing is much less usual, but the skillful and accomplished engraving on one fine Egyptian pot of the predynastic period (*i.e.*, before *c.* 3100 BC) suggests that the practice may have been more frequent than was previously suspected.

Originally, defects of body colour suggested the use of slip, either white or coloured, as a wash over the vessel before firing. A common mode of decoration is to incise a pattern through the slip, revealing the differently coloured body beneath, a technique called sgraffito ("scratched"). Sgraffito ware was produced by Islāmic potters and became common throughout the Middle East. The 18th-century scratched-blue class of English white stoneware is decorated with sgraffito patterns usually touched with blue.

Sgraffito technique

Related to the sgraffito technique is slip carving: the clay body is covered with a thick coating of slip, which is carved out with a knife, leaving a raised design in slip (champlevé technique). Slip carving was done by Islāmic and Chinese potters (Sung dynasty).

Much pierced work—executed by piercing the thrown pot before firing—was done in China during the Ming dynasty (reign of Wan-li). It was sometimes called "demon's work" (*kuei kung*) because of the almost supernatural skill it was supposed to require. English white molded stoneware of the 18th century also has elaborate piercing.

Slip decorating. In addition to sgraffito and carving, slip can be used for painting, trailing, combining, and inlay. The earliest forms of decoration in ancient Egypt, for example, were animal and scenic motifs painted in white slip on a red body; and in the North American Indian cultures coloured slips provided the material for much of the painted freehand decoration.

Slip, too, is sometimes dotted and trailed in much the same way as a confectioner decorates a cake with icing sugar; the English slipwares of the 17th and 18th centuries are typical of this kind of work. Earthenware washed over with a white slip and covered with a colourless glaze is sometimes difficult to distinguish from ware covered with a tin glaze (see below *Decorative glazing*). In consequence it has sometimes been wrongly called faience. The term for French earthenware covered with a transparent glaze (in imitation of Wedgwood's creamware) is *faience fine*, and in Germany it is called *Steingut*. *Mezza-Maiolica* (Italy) and *Halb fayence* (Germany) refer to slip-covered earthenware with incised decoration.

Slip is also used for combed wares. The marbled effect on Chinese pottery of the T'ang dynasty, for example, was sometimes achieved by mingling, with a comb, slips of contrasting colours after they had been put on the pot.

Combed wares and inlay

The Koreans used slip for their inlay technique called *mishima*. The designs were first incised into the clay, and the incisions were then filled with black and white slip.

Burnishing and polishing. When the clay used in early pottery was exceptionally fine, it was sometimes polished or burnished after firing. Such pottery—dating back to 6500 and 2000 BC—has been excavated in Turkey and the Pan-shan cemetery in Manchuria. Most Inca pottery is red polished ware.

Decorative glazing. Early fired earthenware vessels held water, but, because they were still slightly porous, the liquid percolated slowly to the outside where it evaporated, cooling the contents of the vessel. Thus, the porosity of earthenware was, and still is, sometimes an advantage in hot countries, and the principle is utilized in the 20th century in the construction of domestic milk and butter coolers and some food-storage cupboards.

Porosity, however, had many disadvantages—*e.g.*, the vessels could not be used for storing wine or milk. To overcome the porosity, primitive peoples often applied varnishes of one kind or another. Varnished pots were made, for example, in Fiji. The more advanced technique is glazing. The fired object was covered with a finely ground glass powder often suspended in water and was then fired again. During the firing the fine particles covering the surface fused into an amorphous, glasslike layer, sealing the pores of the clay body.

The art of glazing earthenware for decorative as well as practical purposes followed speedily upon its introduction.

On stoneware, hard porcelain, and some soft porcelain, which are fired to the point of vitrification and are therefore nonporous, glazing is used solely for decoration.

Except for tin-glazed wares (see below *Painting*), earthenware glaze was added to the biscuit clay body, which was then fired a second time at a lower temperature. Soft porcelain glaze was always applied in this way. Hard porcelain glaze was usually (and stoneware salt glaze, always) fired at the same time as the raw clay body at the same high temperature.

Kinds of glazes

Basically, there are four principal kinds of glazes: feldspathic, lead, tin, and salt. (Modern technology has produced new glazes that fall into none of these categories while remaining a type of glass.) Feldspathic, lead, and salt glazes are transparent; tin glaze is an opaque white. Hard porcelain takes a feldspathic glaze, soft porcelain usually a kind of lead glaze and can be classified according to the kind of glaze used.

There are two main types of glazed earthenware: the one is covered with a transparent lead glaze, and the other with an opaque white tin glaze.

Tin glaze was no doubt employed in the first place to hide faults of colour in the body, for most clays contain a variable amount of iron that colours the body from buff to dark red. Tin-glazed wares look somewhat as though they have been covered with thick white paint. These wares are often referred to as "tin-enamelled." As noted above, other terms in common use are maiolica, faience, and delft. Unfortunately, these are variously defined by various authorities. The art of tin-glazing was discovered by the Assyrans, who used it to cover courses of decorated brickwork. It was revived in Mesopotamia about the 9th century AD and spread to Moorish Spain, whence it was conveyed to Italy by way of the island of Majorca, or Maiolica. In Italy, tin-glazed earthenware was called maiolica after the place where it was mistakenly thought to have originated. The wares of Italy, particularly those of Faenza, were much prized abroad, and early in the 16th century the technique was imitated in southern France. The term faience, which is applied to French tin-glazed ware, is undoubtedly derived from Faenza. Wares made in Germany, Spain, and Scandinavia are known by the same name. Early in the 17th century a flourishing industry for the manufacture of tin-glazed ware was established at the town of Delft, Holland, and Dutch potters brought the art of tin-glazing to England together with the name of delft, which now applies to ware manufactured in The Netherlands and England. Some misleading uses of these terms include that of applying maiolica to wares made outside Italy but in the Italian style, and faience to Egyptian blue-glazed ware and certain kinds of Near Eastern earthenware.

Although glazed stoneware does not fall into such definite categories as glazed earthenware, to some extent it can be classified according to the kind of glaze used. The fine Chinese stonewares of the Sung dynasty (AD 960–1279) were covered with a glaze made from feldspar, the same vitrifiable material later used in both the body and glaze of porcelain. Stoneware covered with a lead glaze is sometimes seen, but perhaps the majority of extant glazed wares are salt-glazed. In this process a shovelful of common salt (sodium chloride) is thrown into the kiln when the temperature reaches its maximum. The salt splits into its components, the sodium combining with the silica in the clay to form a smear glaze of sodium silicate, the chlorine escaping through the kiln chimney. Salt glazes have a pitted appearance similar to that of orange peel. A little red lead is sometimes added to the salt, which gives the surface an appearance of being glazed by the more usual means.

Crazed and coloured glaze

Some fusion usually occurs between glaze and body, and it is therefore essential that both should shrink by the same proportion and at the same rate on cooling. If there is a discrepancy, the glaze will either develop a network of fine cracks or will peel off altogether. This crazing of the glaze was sometimes deliberately induced as a decorative device by the Chinese.

One method of applying colour to pottery is to add colouring oxides to the glaze itself. Coloured glazes have been widely used on earthenware, stoneware, and porcelain and have led to the development of special techniques in which patterns were incised, or outlined with clay threads (cloisonné technique), so that differently coloured glazes could be used in the same design without intermingling; for example, in the *lakabi* wares of the Middle East.

Unglazed wares

Earthenware, stoneware, and porcelain are all found in unglazed as well as glazed forms. Wares fired without a glaze are called biscuit. Early earthenware pottery, as discussed above, was unglazed and therefore slightly porous. Of the unglazed stonewares, the most familiar are the Chinese Ming dynasty teapots and similar wares from I-hsing in Kiangsu Province, the red stoneware body made at Meissen in Saxony during the first three decades of the 18th century and revived in modern times, and the ornamental basaltes and jaspers made by Josiah Wedgwood and Sons since the 18th century. Biscuit porcelain was introduced in Europe in the 18th century. It was largely confined to figures, most of which were made at the French factories of Vincennes and Sèvres. Unglazed porcelain must be perfect, for the flaws cannot be concealed with glaze or enamel. The fashion for porcelain biscuit was revived in the 19th century and called Parian ware.

Painting. Painted designs are an early development, some remarkably fine work made before 3000 BC coming from excavations at Ur and elsewhere in Mesopotamia, as well as urns from Pan-shan in Manchuria (Northeast Provinces) that date back to 2000 BC.

The earliest pottery colours appear to have been achieved by using slips stained with various metallic oxides (see above *Slip decoration*). At first these were undoubtedly oxides that occurred naturally in the clay; later they were added from other sources. Until the 19th century, when pottery colours began to be manufactured on an industrial scale, the oxides commonly used were those of tin, cobalt, copper, iron, manganese, and antimony. Tin oxide supplied a useful white, which was also used in making tin glaze (see above *Decorative glazing*) and occasionally for painting. Cobalt blue, ranging in colour from grayish blue to pure sapphire, was widely used in the Far East and Europe for blue-and-white porcelain wares. Cupric oxide gives a distinctive series of blues, cuprous oxide a series of greens, and, in the presence of an excess of carbon monoxide (which the Chinese achieved by throwing wet wood into the kiln), cupric oxide yields a bluish red. This particular colour is known as reduced copper, and the kiln is said to have a reducing atmosphere. (For the effect of this atmosphere on the colour of the biscuit body, see above *Firing*). The colours obtained from ferric iron range from pale yellow to black, the most important being a slightly orange red, referred to as iron red. Ferrous iron yields a green that can be seen at its best on Chinese celadon wares. Manganese gives colours varying from the bright red purple similar to permanganate of potash to a dark purplish brown that can be almost black. The aubergine purple of the Chinese was derived from this oxide. Antimony provides an excellent yellow.

Overglaze and underglaze colours

Pottery colours are used in two ways—under the glaze or over it. Overglaze painting is executed on a fired clay body covered with a fired glaze, underglaze painting, on a fired, unglazed body (which includes a body that has been coated with raw or unfired, glaze material).

Earthenware and stoneware are usually decorated with underglaze colours. After the body is manipulated into the desired shape it is fired. It is then painted, coated with glaze, and fired again. The second firing is at a lower temperature than the first, being just sufficient to fuse the glaze. In the case of most tin-glazed wares the fired object was first coated with the tin glaze, then painted, then fired again. The painting needed exceptional skill, since it was executed on the raw glaze and erasures were impossible. The addition of a transparent lead glaze over the painted decoration needed a third firing. In 18th-century Germany especially tin-glazed wares were decorated with colours applied over the fired glaze, as on porcelain. The wares were sometimes called *Fayence-Porcellaine*.

The body and glaze of most hard porcelain are fired in one operation, since the fusion temperature of body and glaze is roughly the same. Underglaze colours are limited

because they must be fired at the same temperature as the body and glaze, which is so high that many colours would "fire away" and disappear. Although the Chinese made some use of copper red, underglaze painting on porcelain is more or less limited to cobalt blue, an extremely stable and reliable colour that yields satisfactory results under both high- and low-temperature firings. On soft porcelain, manganese was sometimes used under the glaze, but examples are rare. All other porcelain colours were painted over the fired glaze and fixed by a second firing that is much lower than the first.

Underglaze pigments are known as high-temperature colours, or colours of the grand feu. Similarly, overglaze colours are known as low-temperature colours, or colours of the petit feu. Other terms for overglaze colours are enamel colours and muffle colours, the latter name being derived from the type of kiln, known as a muffle kiln, in which they are fired. Overglaze colours consist of pigments mixed with glaze material suspended in a medium, such as gum arabic, with an alkaline flux added to lower the melting point below that of the glaze. They were first used in Persia on earthenware (*minai* painting) in the 12th century and perhaps at the same date on Chinese stoneware made at Tz'u-chou.

Lustre painting, gilding, and silvering

Lustre decoration is carried out by applying a colloid suspension of finely powdered gold, silver, platinum, or copper to the glazed and fired object. On a further, gentle firing, gold yields a purplish colour, silver a pale straw colour, platinum retains its natural hue, and copper varies from lemonish yellow to gold and rich brown. Lustre painting was invented by early Islāmic potters.

Pottery may be gilded or silvered. The earliest gilding was done with gold mixed with an oil base. The use of gold ground in honey may be seen on the finest porcelain from Sèvres during the 18th century, as well as on that from Chelsea. Toward the end of the same century gold was applied as an amalgam, the mercury subsequently being volatilized by heating. Silver was used occasionally for the same purposes as gold but with time has nearly always turned black through oxidation.

Transfer printing. The transfer print made from a copper plate was first used in England in the 18th century. In the 20th century transfers from copper plates are in common use for commercial wares, as are lithographic and other processes, such as silk screen printing, which consists of rubbing the colour through a patterned screen of textile material. Combinations of hand-painted and transfer decorations are often used. The outline or other part of the decoration is put on with a transfer print, then parts of the design, such as leaves, flowers, clothing, or water, are painted in.

MARKING

Most porcelain and much earthenware bears marks or devices for the purpose of identification. Stonewares, apart from those of Wedgwood, are not so often marked. Chinese porcelain marks usually record the dynasty and the name of an emperor, but great caution is necessary before accepting them at their face value. The Chinese frequently used the mark of an earlier reign as a sign of veneration for the products of antiquity and, in recent times, for financial gain.

The majority of European factories adopted a device—for example, the well-known crossed swords of Meissen taken from the electoral arms of Saxony, or the royal monogram on Sèvres porcelain—but these, also, cannot be regarded as a guarantee of authenticity. Not only are false marks added to contemporary forgeries but the smaller 18th-century factories often copied the marks of their more august competitors. If 18th-century European porcelain is signed with the artist's name, it generally means that the painting was done outside the factory. Permission to sign factory work was rarely given.

On earthenware, a factory mark is much less usual than on porcelain. Workmen's marks of one kind or another are frequently seen, but signatures are rare. There are a few on Greek vases.

It is often desirable to identify the provenance and the date of manufacture of specimens of pottery as closely as possible. Not only does such information add to the interest of the specimen in question and increase understanding of the pottery art as a whole but it also often throws fresh light on historical questions or the social habits and technical skills of the time it was made. Since ceramics are not affected by any of the agents that attack metal, wood, or textiles, they are often found virtually unchanged after being buried for thousands of years, while other artifacts from the same period are partially or completely destroyed. For this reason archaeologists use pottery extensively; for example, to trace contacts between peoples, since vessels were often widely distributed in course of trade, either by the people who made them or by such maritime nations as the Phoenicians.

Archaeologists' use of pottery

Pottery making is not universal. It is rarely found among nomadic tribes, since potters must live within reach of their raw materials. Moreover, if there are gourds, skins, and similar natural materials that can be made into vessels without trouble, there is no incentive to make pottery. Yet pottery making is one of the most widespread and oldest of the crafts.

Western pottery

ANCIENT NEAR EAST AND EGYPT

In the early 1960s, excavations at a Neolithic settlement at Çatalhüyük, on the Anatolian Plateau of Turkey, revealed a variety of crude, soft earthenware estimated to be approximately 9,000 years old. A more advanced variety of handmade pottery, hardfired and burnished, has proved to be as early as 6500 BC. The use of a red slip covering and molded ornament came a little later.

Handmade pottery has been found at Ur, in Mesopotamia, below the clay termed the Flood deposit. Immediately above the Flood deposit, and therefore dating from a time soon after the Flood (about 3000 BC), was wheelmade decorated pottery of a type usually called Al 'Ubaid. Perhaps the most richly decorated pottery of the Near East, remarkable for its fine painting, comes from Susa (Shushan) in southwest Iran. The motifs are partly geometric, partly stylized but easily recognizable representations of waterfowl and running dogs, usually in friezes. They are generally executed in dark colours on a light ground. Vases, bowls, bowls on feet, and goblets have been found, all dating from about 3200 BC. By 3000 BC pottery was no longer decorated. Earthenware statuettes belong to this period, and a vessel (in the Louvre, Paris) with a long spout based on a copper prototype is the ancestor of many much later variations from this region in both pottery and metal.

Remarkable glazed brick panels have been recovered from the ruins of Khorsabad (Dur Sharrukin), Nimrūd (Calah), Susa, and Babylon. They provide the first instance of the use of tin glaze; although the date of its introduction cannot be certainly determined. A well-known fragment from Nimrūd in the British Museum belongs to about 890 BC, and by the 5th century BC extremely large friezes, one of them about 11 yards (10 metres) long, were being erected at Susa. The presence of lead in the blue glazes derived from copper suggests that the lead may have been added deliberately as a flux, and that this glazing technique, like that of tin-glazing, subsequently was forgotten—to be recovered only at a much later date.

First use of tin glaze

In Egypt, pottery was made in great variety in the predynastic period (up to *c.* 3100 BC), and a hard-fired ware of good quality was attained. The earliest forms of decoration were geometrical or stylized animal or scenic motifs painted in white slip on a red body. There is comparatively little variation until the 26th dynasty (*c.* 664–525 BC), when clay was probably imported from Greece. Most artifacts are vessels of one kind or another, although pottery figures of variable quality were made, some of the later examples (after 500 BC) showing signs of Greek influence.

The so-called faience of Egypt is an unfired ware and thus, strictly speaking, falls outside the definition of pottery used in this article. As early as the 1st dynasty, figures, vases, and tiles of this material were covered with a fired glaze that was coloured turquoise and green with copper oxide. Later, the colouring materials common to

the Egyptian glassmaker, including cobalt and manganese, were added.

ANCIENT AEGEAN AND GREECE

The potter's art first reached the Aegean in the Neolithic, or New Stone Age. All Neolithic vases are handmade, and the best are highly polished; in other respects, the various local schools have little in common, since communications were severely limited in this remote period. The main centres of pottery production lay in Thessaly and Crete. Thessalian potters favoured a red monochrome ware but occasionally attempted simple painted decoration consisting of rectilinear patterns, with a vertical or diagonal emphasis. The Neolithic pottery of Crete is remarkable for its finely burnished surface, any decoration usually incised.

Bronze Age. *Early Bronze Age (c. 3000–2000 BC).* On the mainland, the pottery initiative passed from Thessaly to the Peloponnese and Boeotia. Early Bronze Age pottery from these two areas has been classified into Early, Middle, and Late Helladic, each subdivided into stages I, II, and III. Early Helladic wares show how quickly pottery fell under the influence of the new craft of metalworking: the two leading shapes, the sauceboat and the high-spouted jug, both have metal prototypes. Painted ornament is rare before the final stage (Early Helladic III, or EH III); in the central phase (EH II), the surface is coated with a dark pigment formed from a solution of the clay. This type of paint, later much improved by the Athenians (see below *Attic black-figure and red-figure*), remained the normal medium of decoration on all Aegean pottery until the adoption of a true silicate glaze in Byzantine times.

Influence of metal-working

The contemporary wares of the Cyclades are similar, but more use is made of incised ornament; spirals are common motifs, while some vases bear primitive representations of ships. The pottery of Early Minoan Crete bears simple geometrical patterns, at first in dark paint on a light clay ground (EM I–II), and subsequently in white over a coat of dark paint (EM III). The surface of the ware of Vasiliki in eastern Crete (EM II) has a mottled red and black appearance. The commonest Early Minoan shapes are high-spouted jugs and long-spouted drinking jars resembling teapots.

Middle Bronze Age (c. 2000–1500 BC). After the conquest of the mainland by the first Greeks in the Middle Bronze Age, the local schools of pottery developed on widely different lines. The Minyan ware introduced by the newcomers in an unpainted monochrome body thrown on a fast wheel and fired in a reducing kiln to a uniform gray colour that penetrates the biscuit; the surface is then highly polished and feels soapy to the touch. The shapes are all strongly ridged (carinated) and probably derive from metalwork.

Equally characteristic of this period are the mat-painted wares, which are mainly handmade: here rectilinear patterns are applied in dull black or lilac to a porous white surface. This style, although native to the Cyclades, was also widely imitated on the mainland; in the latest stage the ornament falls increasingly under the influence of the polychrome and curvilinear style of Middle Minoan Crete.

Pottery of Crete

By far the most sophisticated pottery of this epoch was made in Crete, contemporaneously with the first palaces at Knossos and Phaistos. The finest ware (Middle Minoan II) is confined to these two royal capitals and to the Kamáres cave sanctuary whence the style derives its name (Figure 117). Over a dark lustrous ground the ornament is added in red and white, the carefully composed designs striking a subtle balance between curvilinear abstract patterns and stylized motifs derived from plant and marine life. The decoration sometimes takes the form of appliqué molded ornament or barbotine (made of slip) knobs. By the time of MM II the use of the fast wheel had become general, imparting a new crispness to the profiles. Among the commonest shapes are carinated cups (often of eggshell thinness), small, round jars with bridge-spouts, and large storage jars (pithoi). In the course of MM III the fashion for polychrome schemes gradually died out, but at the very end of the period (MM III B) a new naturalistic style was born, inspired by the floral and marine frescoes on

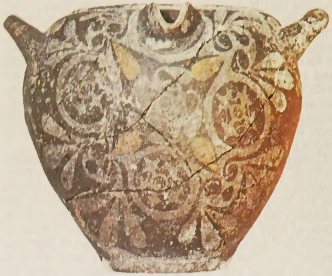


Figure 117: Spouted jar in the polychrome Kamáres style, Middle Minoan, *c.* 1900–1700 BC. In the Archaeological Museum, Iráklion, Crete. Height 24.1 cm.
Hirmer Fotoarchiv, München

the walls of the second palaces. The wide distribution of MM pottery illustrates the vigour of Cretan commercial enterprise; several Minoan emporia were founded in the Aegean Islands, while exports also reached Cyprus, Egypt, and the Levant.

Late Bronze Age (c. 1580–1100 BC). Aegean civilization now reached new heights of prosperity, displayed in the luxurious life of the Minoan palaces and the splendid treasures of the shaft graves at Mycenae. Potters were much influenced by work in richer and more spectacular media: many of their shapes can be traced to originals in gold and bronze found in Cretan palaces and Mycenaean tombs adapted to the shape of the vase.

With the spread of Minoan culture around the shores of the Aegean, Cretan potters exercised a profound influence on the other local schools, and for the first two centuries of this period the vases of the mainland (known as Late Helladic or Mycenaean) are closely related to Minoan models. In the 16th century BC (LM I A), Cretan potters reversed their colour scheme, returning to dark-on-light decoration. Their repertoire includes some abstract motifs (*e.g.*, running spirals and vertical ripples) but is mainly derived from nature, a continuation of the figurative style of MM III B: flowers, grasses, and olive sprays are drawn with charm and spontaneity. After 1500 BC (LM I B) marine creatures are much in evidence, rendered with considerable realism: in a setting of coral and seaweed may be found argonauts, starfish, dolphins, and, above all, the octopus, wrapping his tentacles round the vase. On the palace style amphorae of the late 15th century BC (LM II), however, there is a reaction against this extreme naturalism: plants and marine life continue, but in a more stylized and symmetrical form.

After the destruction of Knossos in *c.* 1400 BC, the artistic initiative passed to Mycenae and remained there until the end of the Bronze Age. In the 14th and 13th centuries BC (LH III A and B), Mycenaen vases were widely exported, not only to Egypt and the Levant but also as far west as Italy and Sicily. In the interests of commerce, pottery was mass-produced, and the Mycenaean colonies on Rhodes and Cyprus were as prolific as the mainland. Some shapes, like the stirrup-vase, were imported for their contents of oil and unguents; others, such as the tall stemmed goblets, were prized for the excellence of their form. Yet, in spite of their high technical standards, the decoration shows a lack of invention. In the absence of any new ideas, the old floral and marine motifs were subjected to an ever-increasing degree of stylization: the flowers degenerate into chevrons and dashes, the octopus into wavy lines. At the same time there is a new tendency to concentrate the decoration into a single focal zone, in anticipation of later Greek pottery. A few large jars bear crude representations of human figures in chariot scenes, probably derived from palace frescoes. (No less schematic are the painted female figurines found in tombs and shrines of this period.) In

the pottery of the 12th century BC, which saw the collapse of Mycenaean civilization (LH III C), there is an abrupt decline in quality as well as in artistic imagination.

Proto-geometric and geometric styles

Early Iron Age. Pottery was the first art to recover its standards after the Dorian invasion and the overthrow of Mycenae. Athens escaped these disasters and in the ensuing dark age became the chief source of ceramic ideas. For a short time Mycenaean motifs survived in debased form but on new shapes. This Submycenaean ware soon gave place to the style known as Protogeometric (*c.* 1100–900 BC) by a natural process of evolution that converted the decaying Mycenaean ornament into regular geometrical patterns; thus, the slovenly spirals were transformed into neat sets of concentric circles, always drawn with a compass fitted with a multiple brush. These circles are the hallmark of Protogeometric decoration, which, like the latest Mycenaean, is confined to the handle zone; in the final stage the rest of the surface is covered with a thick black paint remarkable for its high lustre. Many shapes were inherited from Submycenaean, but all were tautened and vastly improved: the drinking vessels rest on high conical feet, while the closed vases have graceful ovoid bodies. After its invention in Attica, the Protogeometric style spread to other parts of the Aegean world.

Geometric style. In the early 9th century BC Athenian potters introduced the full Geometric style by abandoning circular for rectilinear ornament, the key meander assuming the leading role. At first decoration was restricted to a small reserved area surrounded by the lustrous dark paint; later, as the style approached maturity, more decorated zones were added, until the potter achieved a harmonious balance between light and dark. In the 8th century, after nearly 400 years of abstract decoration, living creatures appear once again, although their style is hardly less angular than the geometric ornament that supports them. Geometric pottery reached its fullest development in the gigantic amphorae and kraters that served as grave monuments in the Athenian Dipylon cemetery; here a funerary scene, showing the corpse on the bier surrounded by mourners, occupies the main panel, while other friezes contain chariot processions, battles on land and sea, rows of animals, and linear geometric designs. The creators of these monumental vases established a continuous tradition of figured painting that persisted on Greek pottery until the end of the Classical period; the immediate consequence of their innovation was a loss of interest in purely abstract design, which became increasingly perfunctory on the latest Geometric vases.

Period of Oriental influence (c. 725–c. 600 BC). After several centuries of isolation, the renewal of contact with the Middle East provided a welcome stimulus to the Greek potter. In art, as well as in commerce, it was Corinth that now led the way. Unlike the Athenians, Corinthian potters specialized in small vases and especially in the tiny aryballos, or scent bottle, which found a ready market throughout the Mediterranean region. There soon arose a style of miniatures that was called Proto-Corinthian; it borrowed much of its repertoire from the fauna and flora of Syrophoenician art. Processions of animals, both real and legendary, are placed in the main friezes, while lotus flowers and palmettes serve as subsidiary ornament. When human beings are depicted, mythical scenes can often be recognized, reflecting the early diffusion of Homeric epic poetry. It was on Proto-Corinthian vases that the technique known as black-figure was first applied: the figures were first drawn in black silhouette and were then marked with incised detail; further touches were added in purple and white.

First use of black-figure technique

Other notable Orientalizing styles arose in Attica, the Cyclades, Laconia, and Rhodes, regional differences in pottery becoming more clearly marked as the Hellenic city-states grew into self-conscious political units. The Athenians still did their best work on large funerary vases. At first they cultivated a wild and grandiose manner in which the figures of men and animals were elaborated in outline; later, incised ornament introduced from Corinth imposed a salutary discipline. Cycladic potters also attempted the grand manner; Laconian work, on the other hand, is confined to a small scale and owes comparatively little to Oriental influence. The Rhodians rarely progressed beyond animal friezes drawn in outline; their style is known as "wild goat", after their favourite quadruped.

Attic black-figure and red-figure. *Archaic period (c. 750–c. 480 BC).* By *c.* 550 BC Athens had once again become the principal centre of pottery manufacture in Greece, having ousted its Corinthian rivals from the overseas markets. Its success is at least partially due to a sudden improvement in technique, for its potters had learned how to obtain the familiar orange-red surface of their vases by mixing a proportion of ruddle, or red ochre, with their clay. As the main medium of decoration, the Athenians perfected a shiny black pigment that was more lustrous than anything that had been hitherto achieved.

In these centuries most of the more important vases were painted either in the black-figure or in the slightly later red-figure technique, so that some explanation of the essential difference is necessary. The red-figure style can be compared with a photographic print, the black-figure with a negative. The latter figures were painted in silhouette in glossy black pigment on the orange-red polished surface. Details were indicated by incised lines and by the occasional use of white and purple, the female figure, especially, being painted in white. Decoration on the red-figure vases was first outlined in black; the surface outside of the outline was then completely covered by the black pigment, leaving the figures reserved in red. Details were added in black, and in dilutions of the black pigment that appear as brown; purple is occasionally found at first but dies out in mature red-figure work. The use of white was revived on the gaudier vases of the 4th century, where yellow brown, gold, and even blue are sometimes used. The forms of Attic black- and red-figure, in the course of centuries, were limited to certain well-defined types, such as the amphora, kylix, krater, and hydria.

Difference between black- and red-figure techniques

The practice of signing vases, already begun in the 7th century, became more common in the 6th. The signatures record either the potter or the painter or in some cases both. The inscription on the celebrated François Vase in the Museo Archeologico in Florence—"Ergotimos made me; Cleitias painted me"—supplies the first positive evidence that, with only occasional exceptions, the two functions had become separate. When the name of a recognizable painter is not known from an inscription, it has become the fashion to name him after the potter with

By courtesy of the trustees of the British Museum



Figure 118: Achilles slaying Penthesilea, Attic amphora in the black-figure style, signed by Exekias, *c.* 530 BC. In the British Museum. Height 41.3 cm.

whom he usually worked: thus, the "Amasis painter" is the habitual colleague of Amasis the potter.

The Attic black-figure style was well developed by the beginning of the 6th century. Among the most favoured subjects were the Labours of Heracles, Theseus, and the revels of Dionysus with his attendant train of satyrs and maenads. The finest Attic black-figure vases were made between 550 and 520 BC, the figures being rendered in a mature Archaic style much influenced by contemporary developments in sculpture. This is the generation of Exekias, the greatest master of the technique (Figure 118). He excelled in painting and in finely engraved detail; he also succeeded, where others had failed, in endowing his figures with mood and emotion, as well as the capacity for action. With Exekias the possibilities of black-figure were virtually exhausted, and after the introduction of red-figure (*c.* 530 BC) it is not surprising that the best artists soon turned to this new technique, which allowed a greater freedom of expression and more naturalistic treatment of the human body. After *c.* 500 BC the only important vases in black-figure are the amphoras presented to victors at the Panathenaic Festival; these have a figure of Athena standing between two pillars and are usually inscribed "I am one of the prizes from Athens."

Early red-figure artists

The early red-figure artists were not slow to exploit the advantages of the new system. Benefitting from the experience of relief sculptors, they had mastered the problems of foreshortening by the end of the 6th century; but since they still avoided any suggestion of depth in their grouping, they were able to convey the illusion of a third dimension without doing violence to the two-dimensional surface of the vase. The most successful work was done in the final years of the Archaic period (*c.* 500–*c.* 480 BC) when the style of the figures, with their formal and elaborate patterns of drapery, was still decorative rather than naturalistic. Monotony was avoided through the use of a wide variety of poses and simple devices for rendering character and mood. Besides the old heroic and Dionysiac themes, many scenes from daily life (especially orgiastic banquets) were now being used (Figure 119).

By courtesy of the trustees of the British Museum



Figure 119: Revelling satyrs, Attic psykter (wine cooler) in the red-figure style, signed by Douris, *c.* 480 BC. In the British Museum. Height 28.6 cm.

Classical period (c. 480–c. 330 BC). This period saw a progressive decline in Attic vase painting. Because of the limitations imposed by the pot surface, the vase painter could no longer keep pace with the rapid advance toward naturalism in the major arts; the occasional attempts at perspective and depth of grouping simply detracted from the shape of the vessel (a mistake repeated in some painting on Italian maiolica in the late 16th century AD). Furthermore, in contrast with the earlier wares, much of the later Attic vase painting shows a saccharine sentimentality and triviality in both the choice of subject and its treatment. Distinguished exceptions are the funerary lekythoi of the late 5th century, decorated in subdued mat colours on a white background. The figures on these vases, isolated and statuesque, share the serenity and restraint of the Parthenon sculptures and suggest something of the grandeur of classical free painting, nearly all of which is now lost. In the 4th century, the figured decoration of pottery had become a degenerate art, and by *c.* 320 BC it had died out in Attica.

In addition to their black- and red-figure vases, the Athenians manufactured plain black-painted wares in great quantity; these follow the shapes of the figured pottery.

Hellenistic period (c. 330–c. 30 BC). After the end of red-figure, Greek pottery is undistinguished. Painted decoration is virtually limited to festoons of ivy, laurel, and a vine in white or yellow over a black ground; the black pigment loses its lustrous sheen and assumes a dull metallic texture. A class of hemispherical bowls, known as Megarian, was made in molds and bears relief decoration in imitation of metal bowls. More remarkable are the contemporary terra-cotta figurines; among the most accomplished are the draped women from Tanagra in Boeotia, whose artistic value is sometimes marred by excessive sentimentality.

ETRUSCAN AND ROMAN

Etruria. At the beginning of the Iron Age (*c.* 900 BC), the most characteristic vessel of the Villanovan culture is the cremation urn. It is usually biconical in shape but sometimes takes the form of a primitive hut, decorated with quasi-architectural ornament in relief.

The first pottery of importance is the Etruscan ware called *bucchero,* which was fired in a reducing kiln. The earliest examples of the 8th century BC, for which the wheel was rarely used, were decorated with incised or engraved geometric patterns. By the 6th century lively and stylized birds and animals were engraved, modelled, or applied in friezes or in conjunction with such geometric patterns as re-entrant (coiling inward) spirals (Figure 120). Later, relief ornament was often executed by rolling a cylinder with design recessed in intaglio over the soft clay, the principle being the same as that used to make Babylonian cylinder seals. Vases with covers in the form of a human head, with arms slipped through fixed ring handles, were made for funerary purposes until about the mid-6th century.

In the late Archaic period the Etruscans excelled in life-size terra-cotta sculptures, of which the outstanding examples are the menacing figure of Apollo, from his temple at Veii, and the large sarcophagi from Caere, with couples of banqueters reclining on the lid. Figures, heads, and busts continued to be produced in the Hellenistic period.

Greek influence

Proto-Corinthian ware was copied with great exactness by Greek colonists as early as 700 BC at Cumae, near Naples. The Etruscans soon learned to use the Greek black pigment, and stylized human and animal figures appear in red, black, and white on a light clay or on the *bucchero* surface. Copies of the black-figure vases were soon so accomplished that it is not always easy to tell exactly where a specimen was made. The red-figure class, however, is rarely difficult to separate from Greek work. The decoration is much more complex and elaborate, and the reverse is often carelessly executed. (Long after the red-figure style had fallen into disuse in Greece, it lingered on in Italy, particularly in the south.)

Roman Empire. The characteristic and most widely dispersed type of pottery of the Roman Empire was the red, polished Arretine ware, so called because manufacture was at first concentrated at Arretium (modern Arezzo). It is sometimes also misleadingly termed Samian ware, from a supposed connection with the island of Samos. The body was generally formed in a mold and was frequently decorated with raised designs. These were achieved by using a mold that had itself been impressed with several stamps arranged in the desired pattern. This decorative technique—which gave the ware yet another name, terra sigillata (clay impressed with designs)—was borrowed from metalwork. The patterns, too, were often influenced by metalwork and



Figure 120: Etruscan amphora of *bucchero* ware decorated with a frieze of horsemen in relief, 6th century BC. In the British Museum. Height 52.1 cm.
By courtesy of the trustees of the British Museum

include floral and foliate motifs, mythological scenes, and scenes from daily life. The potteries at Arretium, which were organized on factory lines, operated between about 30 BC and AD 30. Their products were highly prized and widely exported.

Lead glazing perhaps originated or was rediscovered (the Assyrians having used it) in Egypt. Certainly it was established in the Near East by the 1st century BC. The glazes were generally stained with copper to yield a greenish colour and were sometimes used over relief decoration which, like the designs on Arretine ware, betrays the influence of metalwork. The technique reached Italy and France by the 1st century AD.

Of the other varieties of Roman pottery, lamps made either in a buff or a dark gray clay are common and usually have an impressed or molded design. A few depicting Christian motifs or gladiatorial combats are prized a little more highly than most specimens, but, generally, they are of little value. Molded terra-cotta plaques with reliefs of mythological and other subjects borrowed from Greece were often used to decorate buildings.

ISLĀMIC

In quality, the Islāmic pottery of Syria, Egypt, Mesopotamia, Persia, Afghanistan, and Anatolia rivals even the wares of the Far East, and its influence on the development of European pottery was more profound than that of any other region except China. The Islāmic potter, in his turn, owes an incalculable debt to the Chinese.

Influence of Islāmic pottery on Europe

Near and Middle Eastern pottery was at its best between the 9th and 13th centuries, and its history is closely linked to the fortunes of the caliphate (the dominion of the temporal and spiritual head of Islām). Each dynasty was surrounded in its capital by a wealthy and beauty-loving court that patronized artists and artisans. When one dynasty fell and another established itself elsewhere, it seems that the finest potters emigrated to the new capital, carrying with them their special, and often secret, skills. At first the principal centres of manufacture were Baghdad, al-Fusṭāṭ (old Cairo), and Samarkand; later they shifted to Raqqah on the Euphrates and to Rāy (Rhagae) and Kāshān, both in northern Iran.

Most of the extant pottery has been excavated and consequently is fragmentary. Little made before the 14th century has survived above ground; and tombs, often rich depositories of undamaged wares in other regions of the world, are fruitless because Muslims did not bury pottery with their dead. Only one or two discoveries of undamaged wares have been made: for example, at Gurgan, Iran, entire specimens were found carefully packed in large earthenware jars. They had probably formed part of the stock of merchants, who buried them and fled before the invading Mongols in 1221. Because of deterioration through burial, much Islāmic pottery (like Roman and Near Eastern glass) is iridescent.

Early Islāmic. *Umayyad.* There is little pottery of merit from the period of the Umayyad caliphate (661–750). At this time the capital was at Damascus, and the chief interest of the pottery lies in its mingled Mediterranean and Middle Eastern derivation; for example, attempts were made to synthesize the formal repetitive style derived from the ancient Babylonian and Assyrian civilizations with naturalistic ornament in the Greco-Roman style. When the 'Abbāsids overthrew the Umayyads and moved the capital to Baghdad, the European influence on ornament waned. Good use continued to be made of Western techniques, however, particularly of lead glazes that had been employed by Greek and Roman potters since the 3rd century BC.

'Abbāsid. An event that had a profound effect on the development of the Middle Eastern pottery was the presentation of a number of T'ang porcelain bowls to the caliph Hārūn ar-Rashid about AD 800 (see below *China: T'ang dynasty*). Shortly after this, the first fine pottery was produced in Baghdad and elsewhere in the caliphate. Thus, it seems possible that it was through the example of the Chinese that pottery came to be regarded as an artistic medium instead of a purely utilitarian one. This supposition is borne out by the fact that T'ang wares were in great demand and were imported in large quantities after this date; they and early Islāmic imitations, particularly of the dappled T'ang glazes, have been found in various parts of Mesopotamia and as far apart as Egypt and eastern Persia. Unlike their contemporaries in China, however, Islāmic potters aimed primarily at richness of colour and decoration rather than beautiful shapes and textures. Nearly all their pottery is glazed and is painted with elegant, rather stylized motifs. Floral and foliate ornaments predominate, although complex geometrical patterns are also characteristic. In theory there was a religious ban, formulated in the Ḥadīth (traditions of the Prophet), on all representations of animal life, which were thought to encourage idolatry. In practice, particularly in Persia, the limitation was often disregarded except in the decoration of mosques. The animal figures on pottery are spirited and rhythmical, while the human ones tend to be stiff, resembling those in contemporary miniatures. Arabic calligraphy was commonly and effectively used as an element of design.

Islāmic technical innovations

The Islāmic potters were responsible for a number of important technical innovations, the most influential of which was the rediscovery of tin glaze in the 9th century AD. Though tin was first used by the Assyrians and according to some authorities was discovered as early as 1100 BC, it had fallen into disuse. The 'Abbāsid potters first used it in an attempt to imitate the texture of T'ang wares, but soon it became the vehicle for characteristically Middle Eastern decoration. From Mesopotamia and Persia the technique was later taken to Moorish Spain and then to Italy and other parts of Europe, where it was employed for a number of important wares—maiolica, faience, and delft (see below *European: to the end of the 18th century*).

Like that of tin glazing, the technique of lustre painting was perfected (and probably invented) by Islāmic potters. Again, like tin-glazing, it later passed to Muslim Spain but not to the Far East. Lustre on pottery probably was first used to cover entire vessels, thus simulating vessels made of precious metals that were proscribed by sumptuary laws laid down in the Ḥadīth, which sought to preserve the earlier simplicity of Muslim life. The metallic pigments employed in lustre painting were probably silver and copper in combination, although an occasional ruby glint

suggests that gold may sometimes have been included. After firing, the painting may be dull yellow, golden brown, or olive, tinged with green or red.

Extensive use was made of slip. Wares such as the early Gabrī type of the 11th century and later have a reddish body washed over with white slip. Designs were executed by scratching through the slip to the body beneath (sgraffito). On some later specimens the background was cut away to leave a raised design in white slip or the design was incised through the white slip and then was itself covered with green and brown glazes. The usual motifs are large floral forms, animals, and bold inscriptions. Sgraffito ware became common throughout the Middle East and appears in Egypt and Syria in the 13th century.

Many fragments of Chinese pottery and porcelain have been found at the site of Sāmarrā', on the Tigris, where the 'Abbāsids built their summer palaces in the 9th century (see below *China: T'ang dynasty*). Among the native wares are some made in a buff body decorated in relief under a green glaze; others with monochrome green, white, and yellow glazes or with glazes in imitation of a well-known type of T'ang decoration; and those painted with cobalt blue (perhaps the earliest use of underglaze blue) and further embellished with lustre of various colours.

Sāmānid. To the northeast, beyond the Oxus (modern Amu Darya) River, the Sāmānid dynasty (874–999) became practically independent of the caliphate at Baghdad and fostered a national artistic and literary revival. Sāmānid pottery, which has been found chiefly at Samarkand and Nīshāpūr, differs from the pottery of more westerly regions in technique and style. The best pieces have a reddish body covered with a white, vivid red-brown, or purplish-black slip that was then painted and fired under a lead glaze. The function of the slip, besides providing colour, was to prevent the pigments of the painting from running when the lead glaze was applied. The colours used in painting were the same as those of the slips, with the addition of yellowish green and browns. The designs often consist of the angular Arabic Kūfic characters or stylized birds and floral motifs. The shapes are plain—usually either plates or rather shallow bowls—and the total effect is both bold and elegant.

Egyptian. Egyptian pottery of the Islāmic period was at its best during the Fāṭimid dynasty (969–1171). Wares were at first coarser than those of Mesopotamia because of the poor quality of local materials, and the shapes were less refined, since Chinese influence was absent. Lustre painting (probably introduced in mid-10th century) was nevertheless, excellent in quality. A typical feature is the painting on the backs of dishes, a practice derived from Baghdad and later copied by the Moorish potters of Spain. Signed specimens of lustre ware and tin-glazed wares are known, the best coming from a potter named Sa'd.

Painting on the backs of dishes

Toward the end of the period a much whiter type of ware, with a compact body, came into use and thereafter became common throughout the Middle East. Another widespread group of wares, popular until the 14th century, has decoration carved and incised into the body and is covered with transparent glazes. The patterns suggest the influence of some of the Sung wares of China.

Mesopotamia and Persia. *11th to 15th century.* In the 11th century the Seljuq Turks overran Persia and Mesopotamia, and their ascendancy lasted until the advent of the Mongols during the 13th century. As the Seljuqs had no capital, the most flourishing cities during this time were those on the trade routes. In the 12th century very fine pottery was made in the new white body recently developed in Egypt; it was decorated with bold carving, occasional piercing, and translucent glaze. Most of these wares are said to have been found at Rāy near Teheran, where many other beautiful wares have been excavated. Wares with a sandy body and a clear glaze were painted with a golden-brown lustre, often in conjunction with blue. These seem not to have been made after the city was sacked by Genghis Khan in 1220. Especially associated with Rāy are examples of *minai* painting of uncommon quality. The *minai* technique, a Persian discovery of the 12th century, was a method of decoration in which colours were painted onto a glazed and fired bowl and then fixed by refiring the bowl at a comparatively low temperature. The advantage of the process was that many colours that would not have withstood the heat of the first firing could now be used. The technique may perhaps have influenced the rare examples of overglaze decoration on late Sung or Yüan wares from Tz'u-chou, although it did not come into common use in China until the early part of the 15th century (see below *China: Ming dynasty*).

At Rāy the glaze is cream or turquoise, and the *minai* palette included blue, turquoise, purple, red, green, and white, with the addition of gold leaf. All these colours, except the blue, are mat in appearance, and the style strongly recalls that of Persian manuscript illumination of the 13th century.

Another technique employed at Rāy was the use of silhouette decoration, a kind of sgraffito. The pot was covered with a thick black or blue and black slip, and the design was carved out with a knife. The glazes were applied without colour or stained with copper to yield a brilliant turquoise (Figure 121).

By courtesy of the Victoria and Albert Museum, London; photograph, A.C. Cooper Ltd.



Figure 121: Persian tankard with decoration cut through a black slip and covered with a turquoise glaze, 12th century, found at Soltānābād (modern Arāk, Iran). In the Victoria and Albert Museum. Height 12.7 cm.

Raqqah was a prosperous trading city until it was sacked by the Mongols in 1259. Most of its pottery, which can be dated between the 9th and 14th centuries, is rougher and the designs bolder than those of Rāy. The body is white, inclining to buff, and is covered with a siliceous glaze. Some of the Raqqah fragments are painted with a brownish lustre. Others have designs in relief, sometimes covered with an opaque turquoise glaze or with a bluish-green translucent glaze. In the 12th and early 13th centuries bold designs were executed in black under pale-blue glazes and, more frequently, in blue and black under a clear glaze. Occasionally the glazes were stained purple with manganese.

Kāshān is chiefly famous for its tiles, in fact the words *kāshī* or *kāshānī* ("of Kashan"), are commonly used as synonyms for tile (and have been incorrectly applied to tilework from India). Lustre-painted tiles had been made since at least the 9th century and were used mostly on the walls of mosques and public buildings. Those of Kāshān, particularly in the 13th and 14th centuries, are distinguished by their fine workmanship, brilliance, and intricacy of design. In shape they are square, rectangular, or of interlocking cross or star shapes, each carrying a small part of the total design. The relief inscriptions are frequently picked out with blue pigment.

Kāshān tiles

Also associated with Kāshān are the *lakabi* ("painted") wares made in the 12th century. The term, a misnomer, refers to a variation of the sgraffito silhouette technique mentioned above: an incised design was decorated with different coloured glazes (blue, yellow, purple, and green), which were kept apart by intervening threads of clay. Although a number of *lakabi* wares were also made at Raqqah, the technique was soon abandoned at both places, as the glazes always tended to run out of their compartments during firing, giving a smudged effect.

Both the original site of Sulṭānābād and the nature of the wares that may have been made there are extremely uncertain. Principally associated with it are wares decorated with relief molding under a turquoise or dark-blue glaze or painted in black slip under a clear turquoise glaze. They date from the second half of the 13th century onward. Toward the end of the 12th century the glaze material was frequently mixed with the white-burning clay then in use. In the more highly fired specimens the product is not unlike a primitive soft porcelain, and occasional specimens are slightly translucent. These wares probably inspired the attempts to make porcelain at Florence (see below *European pottery to the end of the 18th century*). Neither stoneware nor true porcelain was ever made in Persia.

After the Mongol conquests of the 13th century the production of pottery practically ceased, except at Kāshān. A slow revival began about 1295, and, although pottery in the Near and Middle East never again reached its former height, some fine wares were made at Sulṭānābād in the 14th century. Good use was made of the rich sombre colours beloved by the Mongols, particularly dark blues, grays, and blacks.

Later Persian. Since the whole of Central Asia now lay under the Mongol domination, overland trade with China greatly increased. By the 15th century Chinese influence, particularly that of Ming blue-and-white, was predominant, and the older styles were tending to die out (see below *China: Ming dynasty*). A group of blue-and-white wares belonging to the 15th and early 16th century are known as Kubachi wares because large numbers of them survived above ground in this town in the Caucasus. They have a very soft body, a brilliant crackled glaze, and rhythmical and spontaneous designs. The later Kubachi blue-and-white is closer to the Chinese originals.

Polychrome appears about 1550, and the palette includes a red related to, though lighter than, the Armenian bole introduced about the same time in Turkey (see below *Turkish pottery*). The best polychrome painting was done on tiles. Tabriz has been suggested as the real centre of manufacture, but although it seems likely that Tabriz was a manufacturing town in view of its tiled mosques and the fact that Tabriz potters were famous abroad (and indeed were either invited or carried off to Turkey on two occasions), no kiln sites have been found there.

One of the later kiln sites in Persia is Kerman, which was the leading pottery centre in the 17th century. Its wares are characterized by a very strong bright blue and a wavy, rather bubbly, glaze. Pseudo-Chinese marks were frequently added to the blue and white. The most usual colours on Kerman polychrome wares are blue, green, browns, and a bright red similar to Armenian bole. The quality of production declined considerably during the 18th century.

Lustre painting

Lustre painting, which had almost ceased in the 13th century, was revived during the second half of the 17th century and perhaps lasted into the 18th century. Its place of manufacture is not known. Most of the objects decorated in this manner are small bottles or spittoons, and their cramped designs are timid and fussy. The lustre is warm brown, often with a strong red tinge, and was sometimes used in conjunction with blue glaze. Another early technique revived at the same time was piercing, formerly practiced in the Seljuq era. There are a number of delicate pierced white wares covered with a colourless glaze, which were imitated in China during the reign of Ch'ien-lung. Pierced pottery and porcelain of this kind was often known in Europe as Gombroon ware, the name of the port (now Bandar 'Abbās) from whence it was shipped.

Chinese celadon was imitated, not very successfully, from the 14th century. In the 16th century other monochrome glazes were produced at Kerman and elsewhere. These and the celadon were frequently decorated with painted or incised ornament—the former a practice quite foreign to Chinese Sung dynasty wares.

During the 18th century most of the pottery produced in Persia was inferior blue-and-white. In the 19th century the standard declined still further with the adoption of the Chinese-inspired *famille rose* palette (see below *China: Ch'ing dynasty*), and only a group of wares made at Teheran between 1860 and 1890 can command any respect. Some excellent peasant pottery with a buff body and lead glaze was made in Turkistan, however.

Syrian. The potters from al-Fusṭāṭ and Raqqah may have migrated to Damascus after their potteries were destroyed by the Mongols, for lustre painting continued in Syria throughout the 13th and 14th centuries after it had ceased elsewhere in the Middle East. The lustre ranges in colour from silver to yellow and dull brown and is often used in conjunction with a blue glaze on big, heavy jars and albarellos (a jar with an incurving waist, used for dry drugs and ointments). Characteristic are gold designs arranged in panels with much use of inscriptions and heraldic devices. The body material is coarse and grayish, and the glaze sometimes has a wide crackle. Lustre painting fell into disuse in Syria about 1400 and might have died out altogether had not the secret meantime been carried from Egypt to Spain (see below *European pottery to the end of the 18th century*). The commonest type of Syrian pottery in the 14th century is a blue-and-black style similar in shape and design to the lustre ware. Rather uncertainly drawn animals appear on some of the vessels.

The earliest known Middle Eastern copies of Chinese blue-and-white were made in Syria at the end of the 14th century. Blue-and-white became commoner on both vessels and tiles in the first half of the next century. Later, the potteries seem to have fallen into disuse until the new mosque built in Damascus by the Turkish ruler Süleyman I (the Magnificent) in the mid-16th century provided a fresh impetus for the industry. The polychrome tiles of the 16th century at first have designs with a hard black outline; later, a more flowing foliate style was developed. A soft purple replaces the Armenian bole of Iznik (see below *Turkish pottery*). Vessels and tiles, gradually declining in quality, continued to be made in Damascus until the end of the 18th century.

Turkish. A branch of the Seljuq Turks occupied Anatolia from 1078 to 1300 and was succeeded by the Ottoman Turks, who first extended their lands westward, conquering Byzantium in 1453 and in the 16th century becoming masters of much of southeastern Europe and the lands lying to the east and south of the Mediterranean. The first notable pottery wares from Turkish lands were the tiles and bricks covered with coloured glazes made in Anatolia for architectural purposes in the 13th century. Mosques in particular were decorated in this way. (Persian influence in decoration suggests the presence of potters from that region.) The art of tilework apparently died out after 1300 and was not reintroduced until about 1415, when Persian craftsmen were brought from Tabriz to decorate the mosques at Bursa and Edirne. Apart from tilework, pottery appears to have received little encouragement until the late 15th century, by which time the chief centre of production was firmly established at Iznik (earlier called Nicaea).

The great era of Turkish pottery

The great era of Turkish pottery (*c.* 1500–*c.* 1580) coincides with the expansion of Ottoman power. Decoration was at first influenced by 15th-century Ming blue-and-white porcelain. The earlier designs were probably taken at second hand from Persian sources, since a distinctly Persian flavour is usually evident. This is indicated by the intricacy of the designs and their arrangement in bands, and by the shapes of some of the vessels, which suggest the influence of metalwork. At one time the wares in this style, which lasted until about 1525, were thought to come from Kūtahya in central Anatolia and are still sometimes known by that name.

At this and later periods the body of Iznik pottery was soft and sandy. It was made from grayish-white clay covered with a thin slip that was usually white, although occasionally red or blue was used as a ground on later wares. Decoration was carried out in underglaze colours under a transparent siliceous glaze. The commonest shapes are flat dishes, but jugs, dishes with a high foot, and bowls are also found. Cylindrical vessels with small rectangular handles set halfway down are flower vases, not tankards, as one might think. A rare form is a pottery version of a mosque lamp.

During the next period (*c.* 1525–50), some wares of which

have been erroneously attributed to Damascus, Iznik pottery was at its finest. Ming blue-and-white was now copied directly; for example, the central motif of grapes on a dish in the Victoria and Albert Museum, London, is an almost exact imitation of a well-known mid-15th century Chinese motif. On the same dish is a characteristic border pattern, which was called the Ammonite scroll border because it was thought to resemble the coiled shell of the fossil ammonite but which is certainly a debased version of the Ming Rock of Ages pattern. This scroll border appears often; a slightly later and even more debased version, which incorporates large S-shaped scrolls, is sometimes known as the dollar pattern.

The palette was gradually expanded to include turquoise, sage green, olive green, purple, and black. Most of the blue and turquoise specimens are painted with flowers. The Chinese flora motifs were almost entirely replaced by tulips, poppies, carnations, roses, and hyacinths in the form of fairly symmetrical sprays springing from a single point. The earliest flowers are often rather more stylized than the later, perhaps because the representation of living things was prohibited by Qur'ānic (Koranic) tradition. Even on comparatively late examples, floral designs are sometimes stylized to the point of abstraction, suggesting that decorators might have suited their patterns to the religious susceptibilities of their customers. An effective abstract pattern is formed from a series of overlapping scales that are usually carefully drawn (Figure 122). The

Scale pattern

By courtesy of the Victoria and Albert Museum, London; photograph, Wilfrid Walter

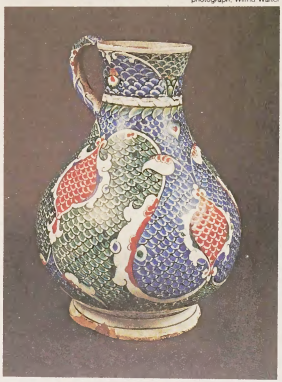


Figure 122: Tin enamelled Turkish jug decorated with the characteristic scale pattern, Iznik (Anatolia), Ottoman period, c. 1575. In the Victoria and Albert Museum, London. Height 27.9 cm.

same ground was later employed in Italy on maiolica and at the Berlin porcelain factory and may have indirectly inspired the series of wares with scale grounds made at Worcester, England.

After about 1550 Iznik pottery enters its third stage. The most notable technical innovation is the use of Armenian bole (sealing-wax red), a thick pigment that stands out in slight relief from the surface of the vessel.

The other great change is that tiles, which had previously been made in small numbers, became all important and remained so until the early 17th century. They were used to provide lavish decoration for the new mosques built at Constantinople by Süleyman I. Once again potters were brought from Tabriz to begin the work. Much use is made of copper green and the new red, the colours very brilliant on the glossy white ground. The tiles, usually square, make up flowing repeating patterns or long high pictures with elaborate borders.

On pottery, symmetrical sprays of flowers continued to be used as decoration until about 1600. Paintings of animals and birds are found occasionally, probably executed by Persian workmen since their resemblance to Persian wares is strong. The rare specimens with human figures were probably painted by Greeks or Armenians for export to the West. Turkish sailing vessels sometimes appear as a decorative motif.

In the 17th century the quality of Iznik wares declined, and by 1800 manufacture had ceased. At Kütahya, pottery making had begun by 1608 and continued into the middle of the 20th century. The wares, though inferior, have some resemblance to those of Iznik with the addition of a yellow pigment.

EUROPEAN: TO THE END OF THE 18TH CENTURY

European wares made before the 19th century fall into six main categories: lead-glazed earthenware, tin-glazed earthenware, stoneware, soft porcelain, hard porcelain, and bone china.

Lead-glazed earthenware was made from medieval times onward and owes little to outside influences. The body is generally reddish buff in colour; the glazes are yellow, brown, purplish, or green. The wares are usually vigorous in form but often crudely finished. Lead-glazed wares fell out of favour when tin glaze became widely known toward the end of the 15th century, but they returned to popularity with the advent of Wedgwood's creamware shortly after the middle of the 18th century. The body of this later lead-glazed earthenware is drab white or cream, the glaze clear and transparent like glass, and the forms precise.

The first important tin-glazed wares came from Italy during the Renaissance, and these colourful examples of the painter's art exerted a profound influence on later work elsewhere. Manufacture spread rapidly, first to France, then to Germany, Holland, England, and Scandinavia. Under the name of maiolica, faience, or delft, it enjoyed immense popularity until the advent of Wedgwood's creamware, after which the fashion for tin-glazed ware declined rapidly.

Stoneware is first commonly seen in Germany during the 16th century; its manufacture was developed in England during the 18th century, culminating in the unglazed ornamental jaspers and basaltes of Wedgwood.

Two other types of ware, less common than those already discussed, are slipware and lustreware. Slip was applied both as a covering over an earthenware body and in the form of decoration, for example on the sgraffito wares of Italy (which owe a good deal to similar wares from Byzantium) and the dotted and trailed slips of 17th- and 18th-century England. Lustre pigments were used in Spain, where they are the principal decoration on the magnificent series of wares referred to as Hispano-Moresque; in Italy, where they supplement other modes of decoration; and much later, in England—although in the last case they are no longer artistically important.

Porcelain and bone china

The manufacture of soft porcelain was essayed in 16th-century Italy under the patronage of Francesco de' Medici, grand duke of Tuscany. Similar attempts were made elsewhere in Italy about the same time, and manufacture is supposed to have been continued at Pisa and at Candiana, near Padua (Padova). The first production of soft porcelain on a considerable scale did not take place, however, until toward the end of the 17th century in France.

In Saxony about 1675 Ehrenfried Walter von Tschirnhaus started experiments to make porcelain from clay mixed with fusible rock. Almost certainly he had made hard porcelain by the end of the century, but manufacture did not become a practical commercial proposition until the year of his death, in 1708. Experiments were continued by his assistant, an alchemist named Johann Friedrich Böttger, who is sometimes credited with von Tschirnhaus' discovery. The factory was established at Meissen about 1710, and the first porcelain sales of any consequence took place at the Leipzig Fair in 1713.

Later, at the end of the 18th century, Josiah Spode the Second added bone ash to the hard porcelain formula to make bone china.

Byzantium. In AD 330 Byzantium became the imperial

capital of the Roman Empire and was renamed Constantinople. The term Byzantine, however, is applied to the period that ended in 1453, when Constantinople was captured by the Ottoman Turks (and renamed Istanbul).

Since it was not a Christian custom to bury pottery with the dead, few wares survive, and chronology is difficult. Most of the surviving wares fall into two classes: one is a red-bodied type, sometimes with stamped relief decoration under a clear glaze; the other, a sgraffito type with human figures, animals, birds, monograms, foliate designs, the Greek cross, and the like, engraved through a white slip and covered with yellow and green glazes. The latter is the commonest type after the 12th century. Both styles were fairly widespread and have been recovered in fragmentary form from excavations at Istanbul, and in Greece, Cyprus, and the Crimea.

Spain. The earthenware of Spain falls into two classes: lustreware and painted tin glazed ware.

Lustreware. The lustre technique spread to Moorish Spain by way of Egypt, but it is impossible to say exactly when it arrived.

Hispano-Moresque pottery

The body of Hispano-Moresque pottery is usually of fairly coarse clay, which has burned to a pinkish buff, covered with a tin glaze containing lead in varying proportions. The lustre, added overglaze, varies in colour from golden to a pale straw, and a coppery lustre almost invariably indicates at least a 17th century date. Many dishes were additionally painted in blue and, less often, with manganese.

Most surviving wares of the early period are dishes of various shapes. Less common are albarellos, wasted drug jars based on a Middle Eastern form (Figure 123). Vases based on the old Iberian Amphora but with two mas-

By courtesy of the trustees of the British Museum



Figure 123. Hispano-Moresque albarello painted with lustre on a blue ground, Valencia, *c.* 1460. In the British Museum. Height 27.6 cm.

sive wing handles (the Alhambra type) are very rare. The decoration on wares of this early period is predominantly Moorish. Fine specimens of this kind are unlikely to be later than 1525. Subsequently, Spanish artists repeated the Moorish designs, but these often degenerate in their hands; and Arabic and the Kūfic script, frequently used by Moorish potters, becomes meaningless. The early designs are, for the most part, plant forms and arabesques, both the vine leaf and the bryony leaf being used. A little later there are magnificently drawn animals in heraldic form, principally lions and eagles. Still later there are deer and antelope, which may owe something to Persian sources. Dishes with coats of arms of noble families surrounded by vine- and bryony-leaf ornament are unusually fine. Many of them were made in Valencia and the neighbouring village of Manises for Italian families. A feature of many of the dishes is the lustre decoration on the reverse. Although often no more than a series of concentric circles, occasionally there are superb eagles and other animals found on dishes from Valencia that are even finer than the obverse designs.

In the 17th century much lustred pottery was made for the cheaper markets and for export to England. The painting is executed in a lustre pigment of deep coppery hue. While this ware is not important in comparison with the early wares, it is often decorative.

Other tin-glazed ware. Although the influence of Valencian lustre pottery on later Italian maiolica is obvious, the wares of Paterna, near Valencia, were hardly less influential in the 14th century. They were decorated in green and manganese, often with motifs taken from Moorish sources; this combination of colours is to be seen in early Italian pottery from Orvieto and elsewhere.

Much tin-glazed pottery of excellent quality was made at Talavera de la Reina, in New Castile, during the 17th and 18th centuries. The palette is characteristic of much Spanish tin-glazed ware; green and manganese play a distinctive part, frequently combined with touches of orange-red and gray. The *istoriato* style of Urbino (see below *Italy*) was copied here, and the Italian painter and engraver Antonio Tempesta (1555–1630) provided a source of inspiration for some of the painting. Alcora, in Valencia, made much faience of excellent quality during the 18th century.

Kinds of tilework

Tilework was particularly common in Spain from the earliest period; according to one proverb, only a really poor man had "a house without tiles." At first tilework was made with a typically Persian technique by which thin slabs of tin-glazed pottery were sawn into pieces and embedded in a kind of mortar (tile mosaic). The *cuerda seca* method of making tiles followed about 1500: outlines were drawn on the surface in manganese mixed with a greasy substance that prevented the coloured glazes used from mingling. Tiles made by the *cuenca* technique had deeply impressed patterns the compartments thus formed being filled with coloured glazes. Tiles were also decorated with lustre pigments.

Porcelain. The early porcelain made at Buen Retiro, near Madrid, in the 1760s, had been justly compared to that of Saint-Cloud. The quality of the ware was good, and some skillful figure modelling was done by Giuseppe Gricci, who had previously worked at Capodimonte (see below *Italy*).

Italy. The pottery of Italy is extremely important not only in itself but for its subsequent influence in other European countries. Indeed, its influence may have spread even farther afield: a few specimens of Ming porcelain have motifs that may have been inspired by it.

There are two well-defined classes of Italian earthenware: maiolica, or tin-glazed ware, and pottery decorated in the sgraffito technique.

Maiolica. Tin-glazing was introduced in the 13th century from the Middle East through the Muslim civilization in southern Spain, wares being shipped from there to Italy by Majorcan traders. The term maiolica was at first applied to this Hispano-Moresque lustreware, but in the 16th century it came to denote all tin-glazed ware.

Italian maiolica is principally noteworthy for its painted decoration, which excelled in technical competence anything produced in Europe since classical times. The painting was executed in several colours on the dry but unfired tin glaze. Great skill was needed, since the surface absorbed the colour as blotting paper absorbs ink, and erasures were therefore impossible. The best wares were given a final coating of clear lead glaze called *coperta*. The range of colours was comparatively limited: cobalt blue, copper green, manganese purple, antimony yellow, and iron red formed the basic palette, while white was provided by the tin-glaze material. When white was used for painting, it was applied onto a bluish-white glaze (*bianco sopra bianco*, or "white on white"), or on a light-blue (*berettino*), or dark-blue ground.

Lustre pigments

Lustre pigments were introduced from Spain. The lustre of Italian wares is often the golden-yellow colour derived from silver, and sometimes it is ruby, suggesting the use of gold. The silver lustre often developed a nacreous effect known as mother-of-pearl (*madre perle*).

The forms of maiolica are few and fairly simple. Generally, they were dictated by the need for a surface on which the painter could exercise his skill; thus, dishes form the greater part of surviving wares. It is doubtful whether most maiolica was ever intended for general use. Dishes were displayed on sideboards and buffets more often than they were placed on the table. Gaily coloured drug jars were a fashionable decoration for pharmacies and include the albarello shape, copied from Spain, for dry drugs, and a spouted jar for wet drugs. Ewers (pitchers) with a trefoil (leafshaped) spout, derived from the Greek oenochoe, were made, as well as the massive jars representative of Florentine work of the 15th century.

The earliest maiolica, beginning in the 13th century, is decorated in green and manganese purple in imitation of the Spanish Paterna ware. Much work of this kind was done at Orvieto, in Umbria, where the characteristic form was a jug with a disproportionately large pouring lip. Orvieto ware has almost become a generic term for anything in this style, although similar vessels were made at Florence, Siena, and elsewhere. It was current in the 14th century and continued in the 15th century, when other colours were added to the palette. The decorative motifs—masks, animals, and foliage—are Gothic, with some traces of Eastern influence.

From Florence came a series of wares painted in a dark, inky, impasto (or very thick) blue. These, too, have Gothic ornament, particularly oak leaves, which came into use sometime before 1450. Heraldic animals also appear on some specimens. This kind of decoration was obviously inspired by Spanish pottery, and a few examples are hardly more than copies. Soon after 1500, Florentine production was concentrated in the castle of Caffagiolo, in Tuscany, and came under the patronage of the Medici family, whose arms appear frequently. A notable addition to the palette here was a bright red pigment, a most difficult colour to attain and one not often used.

Gothic ornament was gradually displaced by classical motifs, such as grotesques, trophies, and the like, which, early in the 16th century, themselves gave way to the *istoriato* style. This style, no doubt inspired by the achievements of contemporary painting, imitates the easel picture closely. Its realism, including the use of perspective, is quite unlike any previous ceramic decoration. The subjects were often classical, but biblical subjects, some taken from the woodcuts of Bernard Salomon (*c.* 1506–61), are frequently represented. Maiolica was often called Raffaelle ware, a tribute to the influence of the painter Raphael (1483–1520), although he, in fact, never made any designs for pottery. In particular the maiolica painters copied his *grotteschi* (grotesques), motifs adapted from those rediscovered in the grottoes of the Golden House of Nero soon after 1500 and so-called in consequence. They are usually fantastic combinations of human, animal, and plant forms. The works of Albrecht Dürer and Andrea Mantegna were also borrowed, often through engravings made by Jacopo Ripanda (Jacopa da Bologna and Marcantonio Raimondi); some examples are almost exact copies, others are freer interpretations. The paintings sometimes occupy the centre of the dish with a border of formal ornament surrounding them, but in many instances, notably those from Urbino, they cover the entire surface. It is often impossible to regard the pottery body as anything more than a support for the painting, its pictorial or narrative subject having been executed with little or no consideration for the nature of the object it decorated. Although pottery decoration is rarely successful unless it is designed to enhance, or at least not to detract from, the shape of the body, an exception must be made for some of these colourful wares: at their best they are highly ornamental.

Istoriato style

The *istoriato* style (Figure 124) probably developed at Faenza (Emilia) in about 1500. One of the earliest and most important centres of production, it had been manufacturing maiolica since before 1450. Almost as early are some examples from Caffagiolo. Castel Durante adopted the same style, and it is particularly associated with the name of Nicola Pelliparo (died *c.* 1542), the greatest of the maiolica painters. He also painted grotesques similar to those of Deruta, in Umbria, which are rather more stylized than the grotesques introduced later in the 16th century at Urbino that are humourous and full of movement. The former are often used as a surround to an interior medallion in the *istoriato* style. Urbino was probably the largest centre for the manufacture of maiolica at the time. The industry there was under the patronage of the Della Rovere family, whose name, meaning "oak tree," led to the adoption of the oak-leaf motif in wreathed form.

Among the early factories, that of Deruta (wich may have been under the patronage of Cesare Borgia) is of considerable importance. Maiolica has been made there from medieval times, and manufacture continues in the mid-20th century. Deruta potters about 1500 were the first to use lustre pigment, which was of a pale-yellow tone,

The Deruta factory

By courtesy of the trustees of the British Museum



Figure 124: "Death of the Virgin," maiolica plaque painted in the *istoriato* style by an unknown Italian artist often referred to as the Master of the Death of the Virgin, Faenza, Italy, *c.* 1510. The composition is taken from Martin Schongauer's engraving of the same subject. In the British Museum. Diameter 26.7 cm.

and they also adopted the Spanish practice of painting designs on the reverse of dishes. They also often covered only the obverse with tin glaze and applied a lead glaze to the reverse—again, a typically Spanish practice. The best work was done before 1540.

The use of lustre pigments at Gubbio, in Umbria, probably started soon after it began at Deruta. The quality of the work was such that maiolica was sent from Castel Durante, Faenza, and even from Deruta itself for this additional embellishment. An interesting series of dishes is that painted with the portraits of young women, often with the addition of a terse and appreciative comment such as "*Bella*," which were made at Deruta, and also apparently at Gubbio.

Maiolica was manufactured in Venice between the 16th and 18th centuries. As might be expected in an important seaport with worldwide trade, its maiolica often shows Eastern influence. The designs of Iznik were sometimes copied (as they were, in fact, on other Italian wares of the period), and imitations of Chinese porcelain of the Ming period gave rise to a style known as *alla porcellana* ("in the manner of porcelain").

Of the later potteries, that of Castelli, near Naples, did excellent work from the 16th century onward, although its later wares tend to become pedestrian. *Istoriato* painting was revived there in the 17th century in a palette paler in tone than that of early work in this style. Much maiolica survives from Savona, in Liguria, a good deal of which is painted in blue in Oriental styles.

Although hardly to be classified as pottery, sculptured reliefs were made by Luca della Robbia (died 1482) in terra cotta and covered with maiolica glazes. He was followed by his nephew, Andrea, and the latter's sons. Giorgio Vasari's suggestion that Luca invented the maiolica glaze is erroneous.

Sgraffito wares. Sgraffito wares are comparatively rare. The technique was derived from Byzantine sources by way of Cyprus which was under Venetian rule from 1472 to 1570. Manufacture was confined to northern Italy, the largest centre being at Bologna. The body of the sgraffito ware was covered with a slip of contrasting colour, the decoration then being scratched through to the body beneath and the whole covered with a lead glaze, which has a yellowish tone. Often the incised designs were first embellished with underglaze colours (blue, green, purple, brown, and yellow) that tended to run during firing. This technique died out finally at the end of the 18th century, but some important work of the kind was done in the late 15th and 16th centuries.

Porcelain. There are only about 50 surviving pieces of the soft-paste porcelain made in Florence at the time of the Medicis, and little is known of its actual production. The earliest definite date for manufacture is 1581. Painting is nearly always in blue with manganese outlines. Most decorative motifs are derived from China, Persia, or Turkey, and the forms usually copy those of Urbino maiolica.

No hard porcelain was made in Italy until Francesco and Giuseppe Vezzi's factory was established in Venice in 1720. It made fine hard porcelain the body of which has a slightly smoky colour. The style is Baroque, and the palette is notable for a brownish red. Another factory, that of Geminiano Cozzi, started in 1764, was the one where most Venetian porcelain was made. Cozzi worked in the Meissen and Sèvres styles and produced some good figures.

Ginori and Capodimonte

The porcelain factory at Doccia, near Florence, was founded by Marchese Carlo Ginori in 1735. Coffeepots in the Baroque style, sometimes painted with coats of arms, are characteristic of the early period. Equally fine figures were made during the 18th century. Porcelain with figure subjects in low relief was made only at Doccia, although it has been repeatedly and erroneously attributed to the soft-porcelain factory established in the royal palace of Capodimonte by Charles III of Naples in 1743. As well as extremely well painted service ware, Capodimonte is renowned for its figures. The factory was transferred to Buen Retiro, near Madrid, in 1759, when Charles became king of Spain (see above *Spain*).

France and Belgium. The medieval pottery of France is difficult to date and classify with accuracy, but lead glaze was in common use by the 13th century at the latest. Proficient sgraffito decoration was done at Beauvaisis (Oise) and at La Chapelle-aux-Pots (Charente-Inférieure).

Palissy rustic ware

Lead-glazed wares of the 16th century. Bernard Palissy began to experiment with coloured glazes about 1539 and, after much difficulty, succeeded in producing his rustic wares in 1548. For the most part these are large dishes made with wavy centres intended to represent a stream, with realistically modelled lizards, snakes, and insects such as dragonflies grouped thereon. They are decorated on the obverse with blue, green, manganese purple, and brown glazes of excellent quality, while the back is covered with a glaze mottled in brown, blue, and purple. Palissy later turned his attention to classical and biblical subjects, which he molded in relief. After his death in 1589, work in his style was continued at the Avon pottery, near Fontainebleau.

Almost contemporary with Palissy's rustic ware is a type of pottery made in the style of the metalwork of the period. It was made at Saint-Porchaire and is sometimes called, erroneously, *Henri Deux* ware, or *faience d'Oiron.* The body is ivory white and covered with a thin glaze. Before firing, designs were impressed into the clay with metal stamps like those used by bookbinders, and the impressions were then filled with slips of contrasting colours. This technique resembles the *mishima* technique of decoration in Korea (see below *Korea*).

Faience, or tin-glazed ware. The technique and the designs of Italian maiolica influenced the development, in the early 16th century, of French faience. There were Italian potters at Lyon in 1512, and, by the end of the 16th century, painting in the manner of Urbino was well established there. Faience was also made at Rouen, probably as early as 1526, and at Nevers toward the end of the 16th century.

A new factory, established at Rouen about 1656 by Edme Poterat, introduced a decoration of *lambrequins,* ornament with a jagged or scalloped outline based on drapery, scrollwork, lacework ornament, and the like. *Lambrequins* were extremely popular and were copied at other porcelain and faience factories. The faience of Nevers, too, is extremely important and shows the Baroque style at its best. In the second half of the 17th century the porcelain of both China and Japan became increasingly well known in Europe, and many designs were borrowed from Chinese sources by potters at Nevers and elsewhere.

The factory of Moustiers in the Basses-Alpes was founded by Pierre Clérissy in 1679. During the early period frequent use was made of the engravings of Antonio Tempesta (1555–1630) as well as biblical scenes. Later came a series of dishes decorated with designs after Jean I Bérain (1637–1711), whose work greatly influenced French decorative art at the time. These designs usually include grotesques, baldacchini (canopies), vases of flowers, and the like, linked together by strapwork in a typically Baroque manner.

In 1709, when Louis XIV and his court melted down their silver to help pay for the War of the Spanish Succession, the nobility looked for a less expensive medium to replace it. In consequence, faience gained in popularity and importance. A great deal was manufactured in the region of Marseilles, the factory of the Veuve Perrin being particularly noted for overglaze painting in the Rococo style. Perhaps the most influential factory was that of Strasbourg, in Alsace (which had officially become part of France in 1697), founded by C.F. Hannong in 1709. The wares—painted in blue, in other faience colours, and in overglaze colours—were much copied elsewhere. Overglaze colours were introduced about 1740, their first recorded use in France. (For the first use in Europe, see below *Germany and Austria.*) Brilliant *indianische Blumen* (flower motifs that were really Japanese in origin but that were thought to be Indian because the decorated porcelain was imported by the East India companies) were painted in a palette that included a carmine similar to the Chinese overglaze rose ("purple of Cassius"). A characteristic copper green was also used. *Deutsche Blumen* ("German flowers") were introduced, perhaps by A.F. von Löwenfinck, about 1750, and inspired similar painting elsewhere. Figures by J.W. Lanz, who also worked in porcelain here and at Frankenthal, are to be seen. Much work was done in the fashionable Rococo style, including objects, such as clock cases and wall cisterns, and tureens in the form of fruit and vegetables. Both faience and porcelain in a variety of decorative forms were used for the banqueting table. Such table decoration, which in the 17th century had been supplied by confectioners who worked in sugar, had become very fashionable in Europe.

Faience and porcelain as table decoration

The wares of Niderviller, in Lorraine, were much influenced by those of Strasbourg. The later figures were probably modelled by the sculptor Charles Gabriel Sauvage, called Lemire (1741–1827), and some were sometimes taken from models by Paul-Louis Cyffle (1724–1806). At Lunéville, not far away, Cyffle worked in a pleasant but sentimental vein and used a semiporcelain biscuit body known as *terre-de-Lorraine,* which was intended to resemble the biscuit porcelain of Sèvres. The work of both Sauvage and Cyffle is extremely skillful.

Faience was made at Tournai (now in Belgium) and at Brussels during the 17th century. Their styles were mainly derivative, but Brussels made some excellent tureens in forms such as poultry, vegetables, and fruits during the Rococo period.

After 1800 most French pottery factories concentrated on the manufacture of *faience fine* (creamware).

Porcelain. In the second half of the 17th century much interest was taken in both faience and porcelain, although the technique of making soft-paste porcelain (*pâte tendre*) had yet to be mastered, and the secret of hard-

paste porcelain manufacture was not discovered until the 18th century.

A factory at Saint-Cloud, founded by Pierre Chicaneau in the 1670s, made faience and a soft-paste porcelain that were yellowish in tone and heavily potted. Much use was made of molded decoration, which included sprigs of prunus blossom copied from the *blanc de Chine* of Tehua (see below *China: Ming dynasty*). Particularly common was a molded pattern of overlapping scales. Most examples are small, but there are some large jardinières (flowerpot holders) that are extremely handsome. The early painted wares were decorated in underglaze blue with typically Baroque patterns, including the *lambrequins* introduced at Rouen (Figure 125). Motifs derived from the designs of

By courtesy of the Museum of Fine Arts, Boston



Figure 125: Porcelain teapot with decoration in underglaze blue, Saint-Cloud, France, *c.* 1720. In the Museum of Fine Arts, Boston. Height 13.9 cm.

Jean Bérain are also to be seen. Polychrome specimens, some of which were decorated in the style of Kakiemon, (see below *Japan: Edo period*), date from about 1730.

At Chantilly, the first soft-paste porcelain was decorated almost entirely in the Kakiemon style, and the body was invariably covered with a tin-glaze. The Japanese period ended about 1740. For some years thereafter simple Meissen styles were copied, in particular the German flowers. In 1753 an edict in support of the newly established factory at Vincennes forbade all other factories to manufacture porcelain or to decorate faience in polychrome; much Chantilly porcelain of the later period, therefore, is creamy white, decorated only with slight flower sprigs in blue underglaze. A transparent glaze was introduced in 1751 and replaced the very unusual practice of covering porcelain with a tin-glaze.

A factory at the Rue de Charonne, in Paris, was started by François Barbin in 1735 and removed to Mennecy in 1748. The early productions were in the manner of Saint-Cloud and Rouen. Later, some excellent flower painting was done, and figure modelling was excellent in quality. Small porcelain boxes from Mennecy, often in the form of animals, are much sought in the 20th century.

Vincennes and Sèvres

The most important of the French factories was established at Vincennes about 1738 and removed to a new building at Sèvres in 1756. Louis XV was a large shareholder in the original company and the factory eventually passed to the crown in 1759. It became state property in 1793, and has so remained.

The factory did not succeed in its attempts to make a practicable soft-paste porcelain until 1745. Much of the work at Vincennes consisted of naturalistic flowers with bronze stalks and leaves, sometimes in vases elaborately mounted in gilt bronze by the court goldsmith, Claude Thomas Duplessis, and others. Meissen was also copied for a short period, but the factory soon evolved its own style, which remained partly dependent on the use of high quality gilt-bronze mounts. A few glazed and painted figures were made, but these gave place in 1751 to figures of biscuit porcelain. In 1757 the sculptor Étienne-Maurice Falconet was appointed to take charge of modelling, a position he retained until 1766. Designs by the painter François Boucher were frequently used by Falconet and others; Boucher's influence is particularly strong during the lifetime of Louis XV's mistress, Mme de Pompadour, who took much interest in the factory. Later, some excellent work in this medium was done by the sculptors Augustin Pajou and Louis-Simon Boizot.

Both at Vincennes and Sèvres much use was made of coloured grounds in conjunction with white panels, which were used for decorative painting of the highest quality (Figure 126). These panels were surrounded by rich and elaborate raised gilding, which was engraved and chased (tooled). The most usual ground colours were a dark underglaze blue (*gros bleu*) and a brighter, overglaze (*bleu de Roi*); also used were turquoise blue, yellow, green, and *rose Pompadour* (often miscalled *rose du Barry* in England).

The porcelain of Sèvres was made to harmonize with the exotic and luxurious style of interior decoration that characterized French court circles. The soft-paste body was of superb quality; and, because the extremely fusible glaze partly remelted in the enamelling kiln, the colours sank into the glaze in a way hardly seen elsewhere.

Hard porcelain of Sèvres

The factory at Sèvres prosecuted the search for the ingredients of hard porcelain with vigour. They were eventually found, after a prolonged search, at Saint-Yrieix-la-Perche, near Limoges, in 1769. The new body was first manufactured soon after 1770, although for a number of years it was only used for biscuit figures. Later, it was employed for dishes and vases decorated in a severe but luxurious classical style. In 1800 the manufacture of soft porcelain was discontinued altogether.

A large number of smaller factories making hard porcelain sprang up, chiefly in and around Paris, in the second half of the 18th century. Some were patronized by members of the royal family, including Louis XVI's wife, Marie Antoinette. A number of provincial factories were also engaged in the same manufacture.

The Tournai factory, in Belgium, which began to make porcelain in 1751, enjoyed the patronage of the empress of Austria, Maria Theresa. Here, and in the associated factory at Saint-Amand-les-Eaux, the work of Sèvres was imitated on a considerable scale.

By courtesy of the Victoria and Albert Museum, London; photograph, A.C. Cooper Ltd



Figure 126: Sèvres vase and cover decorated in reserved panels by Morin, France, 1780. Made for presentation to King Gustav III of Sweden. In the Victoria and Albert Museum, London. Height 49.5 cm.

Germany and Austria. While Germany is principally noted for its superb porcelain, the stoneware of the Rhineland is no less noteworthy. A great deal of faience was also made, though this was less important.

The earliest distinctive type of ware made in markedly Germanic style (*c.* 1350) was the *Hafnergeschirr* ("stove maker vessel"). Originally the term referred to tiles, molded in relief and usually covered with a green glaze, which were built up into the large and elaborate stoves needed to make mid-European winters tolerable. Jugs and other vessels made by these stove makers, however, came to be called Hafner ware by extension when their manufacture began about the mid-16th century. The work of Paul Preuning of Nürnberg is an example of this kind of ware. He decorated his pottery with coloured glazes kept apart by threads of clay (the cloisonné technique). In Silesian Hafner ware, on the other hand, the design is cut out with a knife, the incisions preventing the coloured glazes from mingling. The earliest German stove tiles are lead glazed. Tin glazes came into use about 1500.

Hafner ware

After these beginnings, German pottery developed in two distinct classes: stoneware and tin-glazed earthenware.

Stoneware. The stoneware (*Steinzeug*) came mainly from the Rhineland and, in particular from Cologne, Westerwald, Siegburg, and Raeren (the latter now in Belgium). Manufacture probably began in Cologne about 1540. The body of the stoneware is extremely hard and varies from almost white (Siegburg) to bluish gray (Westerwald); a brown glaze over a drab body is also to be seen (Raeren). The surface is glazed with salt—no more than a smear glaze, pitted slightly, like orange peel. A smooth, though still very thin, glaze was achieved by mixing the salt with red lead. Particularly popular at Cologne in the late 16th century was the "bearded-man jug" (*Bartmannkrug*), a round-bellied jug with the mask of a bearded man applied in relief to the neck. This type was sometimes called a "Bellarmine" in England; the mask was thought to be a satire on the hated Cardinal Robert Bellarmine (Bellarmino), but there is no authority for this assumption. In England, where they were imported in large quantities, they were also known as graybeards. The term tigerware was also used for the mottled brown glaze over a grayish body.

Some of the earliest German stoneware is notable for its remarkably fine relief decoration in the Gothic style. Oak-leaf and vine-leaf motifs were common, as were coats of arms on medallions. The applied relief and stamped decoration was, at times, most elaborate, and the thin glaze lent it additional sharpness and clarity. Reliefs of biblical subjects appear on tall, tapering tankards (*Schnellen*), which were provided with pewter or silver mounts. The *Doppelfrieskrüge* were jugs with two molded friezes (usually portraying classical subjects) around the middle.

Gerald Reitlinger, London, photograph, Wilfred Walter



Figure 127: Gray stoneware jug decorated with cobalt and salt glaze, from the Westerwald region, Germany, early 17th century. English silver mounts are dated 1652. Height 24.1 cm.

They and the tankards were made in Raeren brownware by Jan Emens, surnamed Mennicken, in the last quarter of the 16th century. Emens also worked in the gray body that was used at Raeren at the turn of the century, employing blue pigment to enhance the decoration (Figure 127). At a later date, blue and manganese pigments were used together, and this practice continued throughout the 17th century. Figures were sometimes set in a frame reminiscent of Gothic architectural arcades, and inscriptions of one kind or another are fairly frequent.

The style of the stonewares gradually fell into line with the prevailing Baroque style, particularly toward the end of the 17th century. At Kreussen, in Bavaria, a grayish-red stoneware was covered with a brown glaze, and the molded decoration was often crudely picked out with opaque overglaze colours that had a tin-glazed base. The earliest dated specimen is 1622, which was the first time overglaze colours had been used on pottery in Europe. The technique, learned from Bohemian glass enamellers, was to have some influence in France as well as in Germany.

German stoneware was popular abroad; during the 17th century Sieburg even exported to Japan.

An extremely important type of stoneware was first made shortly before 1710 at a factory at Meissen that was under the patronage of Augustus the Strong, elector of Saxony and king of Poland. It was discovered by E.W. von Tschirnhaus (1651–1708) and J.F. Böttger (1682–1719) during their researches into the secret of porcelain manufacture. It usually varies from red to dark brown and is the hardest substance of its kind known. An almost black variety was termed *Eisenporzellan* ("iron porcelain"), and a black glaze was devised by Böttger to cover specimens of defective colour. Decoration is usually effected by means of applied reliefs, although the black-glazed specimens were sometimes decorated with lacquer colours, as well as with gold and silver. Silvering was not uncommon and was also practiced in other German centres during the early part of the 18th century on both stoneware and porcelain.

Meissen stoneware

A particular feature of Meissen stoneware is the incised decoration done by lapidaries on the engraving wheel. Many specimens were engraved with coats of arms, and grinding into facets (the *Muscheln* pattern) was also practiced. The same methods were used to give a plain surface a high polish. Metal mounts, common Rhenish stoneware, also were sometimes accompanied by insetting precious and semiprecious stones.

Because of the vogue for porcelain, stoneware manufacture declined and was finally abandoned about 1730.

Tin-glazed ware. Faience factories were so numerous that it is only possible to mention the most important of them. Perhaps the earliest tin-glazed wares other than stove tiles are the jugs in the form of owls (with detachable heads to be used as cups) that came from Brixen (Bressanone), in the Tirol. Their shape and style no doubt inspired the later owl and bear jugs made in England during the 18th century. These owl jugs (*Eulenkrüge*) were, at first, used as prizes in archery contests and were sometimes repeated in Rhenish stoneware.

The first manufacture of faience on a considerable scale took place at Nürnberg, and some dishes in the Italian style still survive. Much more is known, however, of the productions of Kreussen, which is chiefly of interest for its blue-and-white faience jugs. The outline of flowers painted in blue is almost cross-sectional in style and terminates in a small spiral—hence the name spiral family.

A factory of Hanau, near Frankfurt am Main, was started in 1661 and remained in operation until 1806. Many of the early wares were decorated with Chinese motifs. A type of jug with a long narrow neck, the *Enghalskrug*, was made in Hanau. Some have a globular body (sometimes copied in China and Japan in blue painted porcelain); others, a spirally fluted body and a twisted handle. Pewter or, less often, silver covers were common. The painting includes coats of arms, landscapes, and biblical subjects. Groups of dots amid strewn flowers (*Streublumen*) are characteristic. Realistically painted German flowers appear shortly before mid-18th century. Most painting is in blue, manganese, and the other less often used faience colours. Overglaze colours do not seem to have been used.

A factory in Frankfurt am Main itself was founded in 1666. Imitations of Chinese motifs as well as biblical subjects were very popular. The blue is brilliant, and the surface usually suggests the use of a transparent overglaze. Narrownecked jugs were commonly made and are sometimes difficult to distinguish from those of Hanau. This centre closed about 1740. At Nürnberg a later factory was established about 1712, continuing until about 1840. Most of the subjects used at Frankfurt and Hanau were repeated at Nürnberg, as well as designs based on the the Rococo engravings of J.E. Nilson (1721–88), which were also popular at many of the porcelain factories. The Rococo style, which spread from France to Germany about the second quarter of the 18th century, is reflected both in the forms and the decoration.

Wares of Bayreuth

The wares of Bayreuth are particularly interesting. Early products were painted with a misty blue, but overglaze colours were speedily adopted. "Leaf and strapwork" (*Laub-und-Bandelwerk*) was a much used type of motif, and excellent work was done by A.F. von Löwenfinck (who is known particularly for his work on porcelain) and Joseph Philipp Danhofer. Perhaps the finest 18th-century faience was made by the factory at Höchst, near Mainz, which also manufactured porcelain. Decoration was usually in overglaze colours, and landscapes, figure subjects, German flowers, and chinoiseries (European delineations of the Chinese scene with a strong element of fantasy) are of a much higher quality than elsewhere. Faience thus decorated with colours applied over the glaze, as on porcelain, was termed *Fayence-Porcellaine* during the 18th century.

An important aspect of both faience and porcelain decoration in Germany is the work of the studio painters, or *Hausmaler,* who brought undecorated faience and porcelain from the factories and painted it at home, firing the decoration in small muffle kilns. For this reason, their work was done in overglaze pigments. At first they mostly used the *Schwarzlot* technique—decoration in a black, linear style that was nearly always based on line engravings. Faience thus decorated dates from about 1660 and is the work of Johann Schaper (died 1670), who had been a Nürnberg glass painter, J.L. Faber, and others. Polychrome enamel decoration was developed by another glass painter, Abraham Helmhack (1654–1724), who mastered the technique as early as 1690, many years before it was adopted by the factories. The more important studio painters are Johann Aufenwerth and Bartholomäus Seuter of Augsburg, J.F. Metszch of Bayreuth, the Bohemians Daniel and Ignaz Preussler, and Ignaz Bottengruber of Breslau. The work of the latter is particularly esteemed.

Toward the end of the 18th century a number of German factories, including some already making faience, made lead-glazed earthenware (*Steingut*) in imitation of Wedgwood, while a factory at Königsberg (now Kaliningrad) imitated Wedgwood's black basaltes body.

Porcelain. The earliest hard porcelain, produced by the factory at Meissen, is smoky in tone, but some improvements were made in 1715 and others in the following decade. Many early specimens were painted with a limited range of overglaze colours of good quality, including a pale violet lustre derived from gold that remained in use until about 1730. In 1720 a painter from Vienna, Johann Gregor Höroldt, was appointed chief painter (*Obermaler*) to the factory; he was responsible for introducing a new and much more brilliant palette, as well as some ground colours (*Fond-Porzellan*). The earliest ground colour to be noted is a coffee brown termed *Kapuzinerbraun,* which was invented by the kilnmaster Samuel Stölzel. The use of blue underglaze proved difficult, and little work of the kind was done. Overglaze painting, on the other hand, was of fine quality and includes topographical subjects, figure subjects based either on harlequin, pierrot, and other characters of the Italian comedy or on the style of the painter Jean-Antoine Watteau and his followers, and flowers in the Oriental style (called *indianische Blumen*) as well as native flowers (*deutsche Blumen*) taken from books of botanical illustrations. A series of harbour scenes from engravings of Italian ports were mostly executed by C.F. Herold (cousin to the *Obermaler*) and J.G. Heintze. Perhaps the most important early wares are the chinoiseries, which appear in great variety. The first work of the kind, much of it painted by the *Hausmaler* Bartholomäus Seuter, is in gold silhouette followed by polychrome painting after designs by the *Obermaler.* The figures are painted in three-quarter length. *Indianische Blumen* motifs were used, and Arita decorations, particularly those of Kakiemon (see below *Japan: Edo period*), were closely copied.

Figure modelling

Little figure modelling was done until about 1727, when the sculptor Johann Gottlob Kirchner was appointed *Modellmeister* and asked to make some colossal figures of animals for the Japanische Palais, the building that housed Augustus the Strong's porcelain collection. Because the medium was unsuited to work of this kind, most of the surviving examples are spectacular and magnificent failures. After the death of Augustus the Strong in 1733 large-scale modelling was practically discontinued, and the new *Modellmeister,* Johann Joachim Kändler, turned his attention to small figures suitable for decorating the dining table.

Assisted by other modellers, Kändler soon made the figures of Meissen fashionable throughout Europe. The first important Rococo work in porcelain appears in Saxony after 1737 when Kändler started to make the Swan service—perhaps the best known of all porcelain services. It is decorated with such motifs as swans, nereides, and tritons. Rococo Meissen was widely sought.

Meissen was the most influential European factory until the beginning of the Seven Years' War in 1756, when it was taken by the Prussians. From then until 1763 it was operated by nominees of Frederick the Great, who virtually looted the factory. By the end of the war, leadership had passed to Sèvres, and the work of Meissen for the next 50 years is much less important than formerly. The transitional Louis XVI style of *c.* 1763–74 is typified by the figure modelling of Michel Victor Acier, who came to the factory to share the position of *Modellmeister* with Kändler in 1764. From 1774 to 1814 the Neoclassical style was increasingly used, and the designs of Sèvres and of Wedgwood (Wedgwoodarbeit) were copied.

Few marks have been so consistently abused as that of the crossed swords of Meissen. Since the 18th century, it has been added to all kinds of unlikely specimens.

The other German factories of the period were, for the most part, established with the aid of runaway workmen from Meissen and Vienna, where Claudius Innocentius du Paquier had started a factory in 1719 with the aid of two men who were themselves from Meissen. Early Vienna hard-porcelain wares are highly prized. Much use was made of leaf and strapwork patterns, and excellent work was done in black monochrome (*Schwarzlot*). The factory passed to the state in 1744, and its later work is competent without being distinguished. Between 1784 and 1805 it became noted for elaborate gilding and coloured grounds, with minutely detailed painting, after Angelica Kauffmann and others, in reserved white medallions.

The Vienna factory provided a number of wandering arcanists (men who possessed the *arcanum,* or "secret," of hard-porcelain manufacture), two of whom helped to establish the Höchst factory, which began manufacture about 1752. This factory is principally noted for excellent figures in the Neoclassical style by Johann Peter Melchior and for the work of Simon Feilner.

A factory in Berlin, started in 1761 and acquired by Frederick in 1763 when he relinquished his hold on Meissen, produced wares with painted decoration of high quality. The decoration made much use of mosaic patterns—detailed diapers (small repeated motifs connecting with one another or growing out of one another with continuously flowing or straight lines) painted over a coloured ground. A large service made in 1819 for presentation to the Duke of Wellington and decorated with scenes from his battles is now in Apsley House, London.

Nymphenburg figures

There is much interest in the figure modelling of Franz Anton Bustelli, who worked at Nymphenburg, a suburb of Munich. The factory, which is still in operation, was started about 1753. Bustelli became *Modellmeister* in 1754 and retained the position until his death in 1763. His magnificent series of figures based on the Italian comedy are the most important expression of Rococo in German

porcelain. The painted wares of the factory were also of fine quality.

Some excellent figures were made at Fürstenberg, where hard porcelain was first manufactured in 1753, and at Frankenthal by such notable modellers as J.W. Lanz, the cousins J.F. and K.G. Lücke, and Konrad Linck. Ludwigsburg, started in 1758, produced porcelain that was grayish in colour and more suitable for figure modelling than for service ware. The figures of artisans by an artist known as the Modeller der Volkstypen (modeller of folktypes) are original and pleasing, and the sculptor, Wilhelm Beyer, did good work in the Neoclassical style.

The Netherlands. During the 17th century, red stoneware was made by Ary de Milde of Delft and others in imitation of the wares of I-hsing (see below *China: Ming dynasty*). Creamware was manufactured at several places at the end of the 18th century. Most Dutch pottery of the period, however, is tin glazed.

Italian potters had settled in Antwerp by 1525, and surviving examples of tin-glazed ware from this period are in the Italian style. Manufacture was concentrated to a great extent in Delft soon after the beginning of the 17th century. By about 1650 the large brewing industry began to decline, and the old buildings were taken over by potters who retained such names as The Three Golden Ash-Barrels, The Four Roman Heroes, and The Double Jug for their potteries. The craftworkers of the town were organized into the Guild of St. Luke, which exercised a considerable amount of control over apprenticeships and established a school of design.

In the 17th century the Dutch East India Company, chartered in 1602, imported Chinese and Japanese wares in great quantities, and the taste for Eastern decoration rapidly ousted Italian fashions. For the greater part of the 17th century decoration was in blue, and Chinese porcelain was closely imitated. In wares of the best quality this imitation is so exact that, without a fairly close inspection, it is possible to mistake them for the originals. Western decorations—biblical and genre scenes, landscapes and seascapes—were carried out in styles similar to Dutch paintings of the period. Tilework was frequently undertaken; many individual tiles have survived, although large panels made up of many tiles are very rarely complete (Figure 128). Blue painting was followed by the use of the usual underglaze faience colours, the outline (known as *trek*) being first drawn with blue or manganese and then filled in. Before firing, the object was covered with an additional transparent lead glaze known as *kwaart*, which made the surface more brilliant. Red was a difficult colour; often when it was to be used, an unpainted space was left during the first firing, and the red was applied afterward and fired at a lower temperature. Gilding is found on the finer specimens and required a further firing. Overglaze colours were introduced by Zacharias Dextra about 1720, and the Chinese *famille rose* patterns were frequently imitated. Among the rarer and more showy examples of delft may be numbered the *Delft dorée*, on which gilding is lavish, and the *Delft noir*, which has a black ground (suggested by Chinese lacquer work) in conjunction with polychrome decoration. Work of this kind is often attributed to Adriaen Pijnacker.

Delft tiles

By courtesy of the Victoria and Albert Museum, London, photograph, A.C. Cooper Ltd.

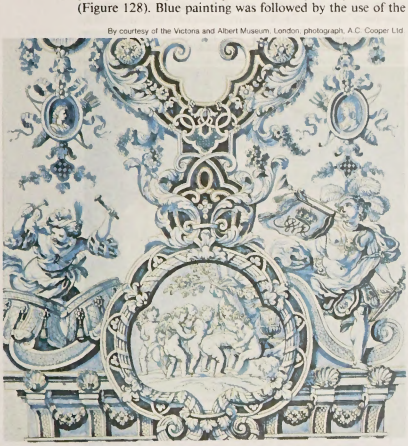


Figure 128: Tin-glazed earthenware wall tile with medallion portraits of William and Mary, Delft, "Greek 'A'," factory of Adrianus Kocks, *c.* 1694. William III of England commissioned these large tiles for Queen Mary's dairy at Hampton Court. In the Victoria and Albert Museum, London. Height 62 cm.

Marks on Dutch delft are extremely unreliable, for many later copies were given the earlier marks of important potteries, especially during the 19th century.

Britain. The medieval pottery of England was affected little by outside influences. Moreover, poor communications prevented the industry from concentrating in any one place; most wares, therefore, are made of local clay by local craftsmen. The potters worked alone or in extremely small groups, and their tools were few and simple. The clay used for the body ranges from buff to red, or, when fired in a reducing atmosphere, from gray to almost black. As with much Japanese pottery, little effort was made to disguise the method by which the vessel was formed, so that pronounced ridges are frequently visible. Both relief and inlaid decoration are found, especially on tiles, and brushed slip was also used to add simple patterns.

Unglazed ware was common, especially in the early period, but a soft lead glaze came into more general use later, the knowledge probably being derived from France. The early glaze varied between yellow and brown according to the iron content of the clay, although a group having a particularly rich brown glaze was made by first washing the pot with slip containing manganese. The use of copper oxide to give a rich green of variable colour dates from the 13th century. During that period, the green, buff, and brown glazes were used in conjunction. Cistercian wares, made in the monasteries before their dissolution in 1536–39, are more precisely finished. They have a dark-brown glaze over a stoneware body and are sometimes decorated with white slip or incised. By far the greatest number of surviving specimens are jugs and vessels for storing liquids; since they have almost always been excavated, a reasonably perfect specimen is a rarity.

Tin-glazed ware. Lead-glazed wares tended to die out after tin glaze reached England via the Netherlands about 1550. At first it was called gallyware, but, with the rise of the Dutch manufacturing centre at Delft, the ware came to be called delft. Its popularity was due to the fact that it could be painted in bright colours. The earliest surviving examples are the Malling jugs, so called because an early specimen of the kind was preserved in the church at West Malling, Kent. These were almost certainly made in London. The colour varies from turquoise to black; a variety with a blue ground flecked with orange was probably suggested by the tigerware from the Rhineland. The jugs usually have silver or pewter mounts. Similar mounts, often of English manufacture, are to be seen on Rhenish jugs imported into England and occasionally on Turkish jugs of about the same period.

Malling jugs

By 1628 a flourishing factory had been established at Southwark, London. Influenced by some Chinese blue-and-white porcelain of the Ming reign of Wan-li (1573–1620), some surviving specimens are decorated in blue, with birds amid floral and foliate motifs. Almost contemporary are some large dishes painted in polychrome colours. The earliest (1600), which is in the London Museum, bears the following couplet: "The rose is red the leaves are grene/God Save Elizabeth our Queene." The dish has a border of blue dashes and is a forerunner of the so-called blue-dash chargers that were popular later in the century. These were decorated with biblical scenes (Adam and Eve being a special favourite), crude portraits of the

kings of England, ships, armorial bearings, and the like. The influence of Italian maiolica and Chinese porcelain can be seen in the border designs.

Many wine bottles are extant, often with the name of the wine (Sack, Claret, etc.) painted in blue and a date. Others are more elaborately decorated, a few are in polychrome.

Toward the end of the 17th century service ware became more frequent (although tea ware was now scarce). Blue-and-white was still made in large quantities, but a polychrome palette was more in evidence, and the influence of Dutch potters is often obvious.

Chinese influence, which had been particularly strong in the early part of the 18th century, tended to persist, particularly at Bristol. The Rococo style was used to a limited extent. Later, some not very successful attempts were made to utilize The Neoclassical forms. Overglaze colours on tin-glazed wares appear after mid-century. These colours, a special pallet now called Fazackerly colours, were probably used only at Liverpool.

Fazackerly colours

The main centres of production of tin-glazed ware were in London (Southwark and Lambeth), Bristol, and Liverpool, although there were smaller potteries elsewhere. One of them—Wincanton in Somerset—made frequent use of manganese, which produces purple and purplish-black colours. The tin glaze fell into disuse about the turn of the 18th century, its place having been taken by Wedgwood's creamware. (In the mid-20th century manufacture has been successfully revived at Rye, Sussex.)

17th-century slipware. Wares decorated with dotted and trailed slip were made at Wrotham, Kent, and in London during the first half of the 17th century. Wrotham is noted principally for drinking mugs with two or more handles, known as tygs; and London for dishes with such pious exhortations as "Fast and Pray," obviously inspired by the Puritans. Manufacture was also started in Staffordshire, and many surviving examples were signed by the potter in slip. The work of Thomas Toft is particularly valued. The best work of this kind was done before the end of the 17th century, and although it may fairly be described as peasant ware, many of the earlier specimens are vigorously decorated and amusing. Manufacture continued until the end of the 18th century.

Stoneware. The popularity of Rhineland stonewares in England, as well as that of the newly imported Chinese stoneware teapots from I-hsing kilns (see below *China: Ming dynasty*), led to attempts to imitate both kinds. The first patent for making copies of porcelain and Cologne ware known to have been exercised was awarded to John Dwight (*c.* 1637–1703) of Fulham in 1671. In addition to German stoneware, he made a brown-glazed stoneware decorated with stamped ornament that was continued at Fulham after his death and has been extensively reproduced since. He probably never made any porcelain, but he mentions red china, which can only refer to imitations of the I-hsing stoneware.

The brothers John Philip and David Elers, of German origin, made red stoneware at a factory in Staffordshire. It is difficult to separate their work from that of Dwight (at Fulham), on the one hand, and that of their Staffordshire imitators, on the other. Most wares are decorated with stamped reliefs, the Chinese prunus blossom being comparatively common. The tendency to utilize patterns from silverwork, which is apparent on some examples, may be connected with the fact that the Elers had been silversmiths. The Elers' migration to Staffordshire perhaps can be regarded as the starting point for the large modern industry that has grown up in that area. Certainly from this time onward Staffordshire wares tend to lose their peasant character and to approach a factory-made precision that was to be general by the end of the 18th century.

Salt-glazed stoneware

The earlier red- and brown-glazed stonewares were replaced about 1690 by a salt-glazed stoneware that was regarded as an acceptable substitute for porcelain. It varies in colour from drab to off-white, the glaze on later specimens often having a richer, more glassy appearance due to the addition of red lead to the salt. One of the earliest varieties is decorated with reliefs stamped from pads of clay that were applied to the surface.

18th-century developments. The "scratched-blue" class of white stoneware dates from about 1730 and is decorated with incised patterns, usually touched with blue. Decoration is floral, and inscriptions and dates are fairly frequent. Its manufacture continued until about 1775.

From the 1730s molded patterns in relief were popular, the clay being pressed into molds of metal, wood, or fired clay. The introduction of plaster of paris molds around 1745 gave much greater scope and led to the development of intricate shapes in the finer varieties of white stoneware. The patterns greatly increased in sharpness and elaborate piercing is to be seen.

Transfer printing was first used about 1755, possibly at Liverpool, which produced wares of all kinds, including tiles, using this decorative technique.

The earliest use of overglaze colours belongs to the same period—previously, white wares had been sent to Holland for decoration. The Englishman who first mastered the technique was William Duesbury. Established as a decorator in London by 1751, he concentrated on painting porcelain, but he also seems to have overglaze-painted stoneware from Staffordshire. Some extant brilliantly painted figures are probably from his studio. A little earlier than Duesbury's overglaze-painted figures are the uncoloured pew groups, which consist of two or three figures seated on a high-backed settle or pew, modelled in a primitive and amusing fashion. A rich blue overglaze ground, often called Littler's blue after William Littler, who is thought to have invented it, was much used on the salt-glazed stoneware, as well as the porcelain, made at Longton Hall, a factory that operated in Staffordshire from about 1750 to 1760 and that was also associated with Sittler.

John Astbury is particularly associated with a type of brown-glazed ware decorated with stamped pads of white clay. Some of the earliest Staffordshire figures in brown and white clay covered with a lead glaze have been attributed to Astbury.

Thomas Whieldon (1719–95) of Fenton Low, Staffordshire, manufactured agateware—that is, ware made by combining differently coloured clays or by combing together different colours of slip. In the former method the clays were usually laid in slabs, one on the other, and beaten out to form a homogeneous mass in which the colours were inextricably mingled. Agatewares seem to have been made in Staffordshire between 1725 and 1750, the earlier specimens being salt glazed, while the later ones were covered with a colourless lead glaze. Whieldon

Agateware

By courtesy of the Victoria and Albert Museum, London; photograph, Wilfrid Wallor



Figure 129: Mounted Hudibras, creamware decorated with coloured glazes by Ralph Wood, Staffordshire, *c.* 1765. In the Victoria and Albert Museum, London. Height 29.8 cm.

is most famous for his use of coloured glazes that were mingled to give a clouded or tortoiseshell effect and were used on an earthenware body, sometimes over molded decoration. A few naïvely modelled figures with this type of glaze are attributed to him. From 1754 to 1759 he was in partnership with Josiah Wedgwood, who developed the fine green and yellow glazes to decorate molded wares in the form of pineapples, cauliflowers, and the like.

Coloured glazes were also used by Ralph Wood I (1715–72) of Burslem, Staffordshire, for decorating an excellently modelled series of figures in a creamware (lead-glazed earthenware) body, the finest, perhaps, a mounted Hudibras (Figure 129) in the Victoria and Albert Museum. Many of these figures are attributed to the modeller Jean Voyez, who was much influenced by the work of Paul-Louis Cyfflé at Lunéville (see above *France and Belgium*). Ralph Wood I is also noted for the typical English Toby jug (first made soon after 1700), which is a beer jug in the form of a man, usually seated and holding a pipe and a mug, the hat (where present) forming a detachable lid. Very popular, it continued in production for many years. Enoch Wood, another member of the family, joined Ralph Wood II in partnership as Enoch Wood & Co., which lasted until 1790. They made most of the wares current in Staffordshire at the time, as well as some excellent figures decorated with overglaze colours.

Wedgwood

Josiah Wedgwood (1730–95), the most famous of all the Staffordshire potters and the most important exponent of Neoclassicism in the field of pottery, is celebrated chiefly for his fine jasper and black basaltes stonewares, but his creamware was undoubtedly the more influential in the 18th century. It was well finished and clean in appearance, with simple decoration in good taste, often in the popular Neoclassical style. His wares appealed particularly to the rising bourgeois class, both in England and abroad, and porcelain and faience factories suffered severely from competition with him. Surviving factories switched to the manufacture of creamware (*faience fine* or *faience anglaise*), and the use of tin glaze almost died out.

Wedgwood secured the patronage of Queen Charlotte (wife of King George III) for his creamware in 1765 and renamed it Queen's ware. Much of it was transfer printed by John Sadler and Guy Green at Liverpool. Evidence of its popularity and importance is provided by the enormous service of 952 pieces made for Catherine the Great's palace of La Grenouillière, in St. Petersburg.

The basaltes ware, also called black porcelain or Egyptian ware, was a type of stoneware introduced about 1768. Like the jasper that followed, it was used almost entirely for ornamental work—vases, ewers, candlesticks, plaques, medallions, and tea and coffee ware. Some of it was painted in what Wedgwood called encaustic enamel in imitation of Greek red- and black-figure vases, but most of the decoration was either molded and applied or incised by turning on a lathe.

Jasperware

Jasper, introduced about 1775, is a fine-grained white unglazed stoneware, slightly translucent when thinly potted or fired above the normal temperature. Undoubtedly it was inspired by the biscuit porcelain of Sèvres. Its name derives from the fact that it resembles the natural stone in hardness. At first the body was stained blue (with applied decoration in white). Other colours, such as sage green, lilac, black, and yellow, followed speedily. Like basaltes, jasper was used mainly for ornamental wares, but perhaps the most interesting products are the portrait medallions of contemporary notables. Vases do not appear to have been made until after 1780 (Figure 130). In 1790 Wedgwood produced the first copies of the Portland vase, a magnificent Roman cameo glass vase of dark blue glass decorated with white figures, at that time owned by the Duke of Portland but now in the British Museum. The vase was reproduced in later years, particularly in Victorian times both by Wedgwood in jasper and by Northwood in glass. Wedgwood's jasperwares were imitated in biscuit porcelain at Sèvres, and Meissen produced a glazed version called *Wedgwoodarbeiten*. Less influential was the red stoneware (*rosso antico*), which sometimes had an enamelled decoration of classical subjects, and caneware, a buff stoneware.

Lustre pigments introduced into England toward the end



Figure 130: Jasperware vase molded with "the crowning of a kitarist" in white relief against a pale blue background, impressed Wedgwood, 1786. In the British Museum. Height 43.2 cm.
By courtesy of the trustees of the British Museum

of the 18th century were used in a manner quite different from the earlier styles of other countries (see above *Spain* and *Islāmic*). To simulate silverwork, wares were completely covered with platinum lustre, which remains unchanged in colour after firing (silver itself yields a pale straw colour); the amount of metal used was extremely small. Such wares were known as poor man's silver. Wares were also painted or stencilled with lustre patterns. The most valuable type commercially were the resist lustres, which have a lustred background and the pattern reserved in white. They were made by painting or stencilling the pattern on the glaze with shellac, which resisted the subsequent application of the metallic pigment. Silver lustre was rarely used, but gold lustre, which gives variable colours from pink to purple, was fairly common. (Copper, the colour of which remains more or less unchanged in its lustre form, was used throughout the 19th century for common wares.)

Poor man's silver

Porcelain. A factory for porcelain manufacture, using a soft-paste body similar to that of Saint-Cloud, was established in Chelsea, London, about 1743 by Charles Gouyn and Nicolas Sprimont, the latter a silversmith. The rare surviving specimens include jugs molded in the form of goats and further decorated with an applied bee, obviously based on a silver prototype that no longer exists. (Extant examples of the latter are 19th-century forgeries.) These goat and bee jugs are often marked with an incised triangle, which was then the mark in use. About 1750 a new body was adopted, together with the familiar mark of an anchor, which was raised on a small medallion until about 1752, painted in red until about 1756, and executed in gold thereafter. The work of the Chelsea factory was extensively influenced by Meissen until about 1756, the styles of Sèvres superseding it in the gold-anchor period. Wares marked with either the raised or the red anchor are the most highly valued; the painting of these is excellent in quality. Some of the best wares were painted by an Irish miniaturist, Jeffrey Hamet O'Neal. The gold-anchor-marked wares are noted for rich gilding and some fine coloured grounds that, on occasion, rivalled those of Sèvres (Figure 131). The figures in the later Rococo style are generally inferior to those of the earlier red-anchor period. Some Chelsea porcelain from 1760 onward was painted in the studio of James Giles of Clerkenwell. The factory was bought by William Duesbury of Derby (see below) in 1770

Artists' marks determine value



Figure 131: "The Music Lesson," enamelled soft-paste porcelain, English, Chelsea, gold anchor period, c. 1765. In the Metropolitan Museum of Art, New York. Height 38.1 cm.
By courtesy of the Metropolitan Museum of Art, New York, collection of Irwin Untermyer

and entered a phase known as the Chelsea-Derby period. The Neoclassical style was introduced together with the figure in biscuit porcelain made fashionable by Sèvres. It closed finally in 1784.

A group of figures, the best known examples of which are those portraying a girl in a swing, were made in the 1750s—possibly at Chelsea but more probably at a short-lived factory staffed by workmen who had seceded from Chelsea. A class of figures characterized by an apparent retraction of the glaze from the base—dry-edged figures—are attributed to a factory established at Derby about 1750. This enterprise apparently petered out and another factory in Derby was started in 1756 by Duesbury (who was later to buy the Chelsea factory). It advertised itself as the second Dresden and is noted toward the end of the century for the excellence of its painting by Zachariah Boreman, William Billingsley, and others.

Longton Hall in Staffordshire made figures and a good deal of service ware molded in the form of leaves. A rich blue ground (Littler's blue) was used on porcelain and salt-glazed wares alike. Its wares are rare and much sought.

The Bow factory (London) was started as early as 1744 with the aid of clay brought from Virginia by the American settler Andrew (André) Duché, who had discovered the secret of manufacture quite independently some years before (see below *Colonial America*). An amusing and primitive class of Bow figures was executed by an anonymous artist known as the Muses Modeller, because the most typical figures portray the Muses. Generally speaking, Bow wares are unsophisticated, and the factory obviously catered to prosperous tradesmen, a market ignored by Chelsea. An important technical innovation took place at Bow in 1750, when calcined bones were added to the porcelain body. This was the first major departure from the French soft-porcelain formula, which was fundamentally a mixture of clay and ground glass. Bone ash was added to soft porcelain by Chelsea about 1755, by Lowestoft (which mainly copied Bow styles) in 1758, and by Duesbury to Derby porcelain in 1770, when he purchased the Chelsea factory.

Development of English bone china

About 1800 at his factory at Stoke-upon-Trent, Staffordshire, Josiah Spode the Second added calcined bones to the hard-porcelain formula to produce the standard English bone-china body (see below *19th century*).

Another variation on the original soft-porcelain body was introduced at a factory in Bristol started by Benjamin Lund about 1748. Clay was mixed with a fusible rock called steatite (hydrous magnesium silicate), the principle being similar to that used in the manufacture of hard porcelain. This factory was transfered to Worcester in 1752 and still manufactures fine porcelain. In the 18th century, scale grounds, which consisted of patterns of overlapping scales in various colours, were particularly popular. Transfers taken from engraved plates were also extensively used for decoration. After 1783 wares show a progressive decline in taste. A second factory was established at Worcester by Robert Chamberlain in 1786 (see below *19th century*).

William Cookworthy discovered the secret of hard porcelain independently after many years of experiment. In 1768 he opened a factory at Plymouth (which was transferred to Bristol in 1770) that made figures in the style of Bow and Longton Hall. Richard Champion acquired the patent for hard porcelain in 1772 and manufactured tableware Neoclassical in style and excellent in quality. The patent was bought by a syndicate that established a factory at New Hall, Staffordshire, in 1782 and made a humble variety of wares for about 40 years.

Scandinavia. The faience industry spread to Scandinavia mainly because of migratory workmen from Germany. A number of factories in Denmark, Norway, and Sweden during the 18th century made faience and creamware in the English manner. A distinctive Scandinavian production was that of bowls, made in the shape of a mitre, for a kind of punch called bishop. The most important factories are those of Rörstrand and Marieberg (Koja) in Sweden. A typical Rococo concept to come from Marieberg is a vase standing at the top of a winding flight of steps. Called a terrace vase it is often decorated with a rabbit or some other animal.

In 1774 a factory at Copenhagen directed by Louis Fournier, a modeller from Vincennes and Chantilly, began the manufacture of true porcelain. The factory was acquired in 1779 by King Christian VII of Denmark and Norway. In 1789 the factory started work on an enormous service, originally intended for Catherine the Great, each piece of which was painted with a detailed picture of a Danish flower. This service, the "Flora Danica," is now in Rosenborg Palace, Copenhagen. Numerous skillfully made figures were also produced. The factory continues to produce fine porcelain.

Service for Catherine the Great

Switzerland and Russia. A factory started near Zürich in 1763 and directed by Adam Spengler made both faience and porcelain and, after 1790, creamware. Delicate figures, some modelled by J.V. Sonnenschein from Ludwigsburg, and good-quality service ware were produced.

The factory of St. Petersburg was established about 1745. Later production was on a fairly large scale, and the work of Sèvres and Meissen was freely copied. Some good original work was also done, and well-modelled figures of Russian peasants were made toward the end of the century. Even better figures were made at a factory in Moscow founded about 1765 by an Englishman named Francis Gardner. Many factories at Moscow and elsewhere in Russia were established during the 19th century.

Colonial America. There is little detailed information about the pottery made by the early European settlers in North America. Most of it was manufactured locally for local needs and from the clays that were nearest to hand. Since most of these contained iron in varying quantities, the pottery body burned to colours between buff and red. Until kilns capable of reaching a high temperature were constructed, manufacture was limited to earthenware. Lead glazes were commonly used. Slips, both as a wash and as trailed decoration, were employed, and sgraffito decoration is known. Most of this pottery was made for practical rather than decorative purposes. A few potteries were established in the 17th century in Virginia, Massachusetts, and New Jersey; and in eastern Pennsylvania, German settlers started work as early as 1735 making slip-painted and sgraffito earthenware in their own traditions.

Perhaps the most important development in colonial America took place in Savannah, Georgia, where Andrew Duché started a pottery about 1730. He interested himself in the manufacture of porcelain and discovered the china clay and feldspathic rock necessary to its manufacture. By

1741 he appears to have made a successful true porcelain but failed to gain adequate financial assistance to develop it. He therefore travelled to London, arriving in 1744, and tried to sell the secret to the founders of the Bow factory in London. Their interest is certain, since the patent specification subsequently filed specifically mentions *unaker,* said to be the Cherokee name for china clay. Duché returned to Virginia by way of Plymouth and there spoke with William Cookworthy, later to be the first manufacturer of true porcelain in England. It is still not known to what extent Duché actually manufactured porcelain; but since the *Bristol Journal* for November 24, 1764, refers to the import of some specimens of porcelain said to have been made in Georgia, there is little doubt that the first porcelain to be made in an English-speaking country came from North America. The Cherokee clay was shipped to England from time to time during the 18th century. Wedgwood imported several tons of it to use in the development of the jasper body.

Hard porcelain in America

By 1765 potteries were being established on a sufficient scale to warrant an attempt to recruit workmen from Staffordshire. Wedgwood wrote at the time: "They had a agent amongst us hiring a number of our hands for establishment of new Pottworks in South Carolina."

The manufacture of tin-glazed ware began in Mexico soon after the Spanish Conquest in the first half of the 16th century. Spanish styles predominated, especially that of Talavera, but Chinese influence occurs in the 18th century. The wares became a kind of inspired folk pottery in the 19th century.

19TH CENTURY

There is a fundamental difference between work done before the Industrial Revolution, the effect of which began to be felt in the pottery industry before 1800, and that done subsequently. A student of the older wares, particularly those of the East, may find much of the later work difficult to accept because of its machine finish. When an object is made by hand it is never exactly the same as any other object, nor are the processes by which it has been formed and decorated disguised. Consider, for example, a Sung dynasty pot or a specimen of Japanese tea ceremony ware, whose imperfections of finish by factory standards are an integral part of their beauty and character, or the glaze of a *Kuan* vase, which would lose its individuality if it possessed the smooth finish of a factory-made specimen. The technical precision of the 19th century, which made its products indistinguishable from one another, and the careful concealment of the means by which the end had been achieved, were both unprecedented and deleterious. Style and craftsmanship degenerated steadily in the factories. The situation was aggravated by the Great Exhibition of 1851, which encouraged manufacturers throughout Europe to vie with each other in producing wares displaying virtuosity unhampered by questions of taste; for example, from as far afield as St. Petersburg, theretofore outside the mainstream of European development, came some particularly colossal and hideous vases in a debased Neoclassical style—which were described by a contemporary writer as "second to few of the productions of Dresden and Sèvres for beauty of outline and perfection of finish."

The degeneration of taste

Those who bought these wares—as well as those who produced them—contributed to the degeneration of taste. Before the advent of mass communications in the 20th century, new fashions originated in the wealthiest stratum of society (which was usually also the most cultivated) and filtered downward. As a result of the political and economic effects of the Seven Years' War (1756–63), combined with the beginning of the Industrial Revolution, the European bourgeoisie prospered, and their wealth enabled them to become patrons and arbiters of taste. Primarily interested in the arts as a means of display or as status symbols, they demanded an excess of intricate and expensive ornament. In East Asia the same process of degeneration began at the same time, at least partly as a result of the large number of export orders received. This pernicious influence was kept at bay for awhile by the emperor Ch'ien-lung, who stigmatized the English as cultural barbarians, but became more pronounced in the 19th century. Similar tendencies may be seen in Japanese pottery after 1853, when many factories worked almost entirely in styles demanded by their customers in the West.

Britain. *Porcelain.* The Neoclassical style, which had been popular during the middle years of the 18th century, gradually lost its earlier simplicity. In France, the rise of Napoleon brought in its train the ostentatious Empire style (copied, for the most part, from the decorative art of imperial Rome), which had much influence in England during the Regency period (1811–20). It is noticeable on the porcelain vases made at such factories as Worcester, Derby, and Rockingham. They were often decorated with well-painted topographical subjects that were no longer confined by frames but ran around the vase as a continuous landscape. Flower painting was often of excellent quality and was much influenced by the work of William Billingsley, a flower painter who worked at Derby toward the end of the 18th century.

At Worcester a factory established by Robert Chamberlain in 1786 produced porcelain decorated in a debased Japanese style. Because of their gaudy colour—iron red and underglaze blue coupled with lavish gilding—some Japan patterns are called thunder-and-lightning patterns. Similar Japan patterns were being employed at Derby and at an older Worcester factory, although much of the work of the latter was more restrained. Some of the best painting at the old factory was executed by Thomas Baxter, who used marine shells as a subject.

It has been said, unfairly, that Josiah Wedgwood, by developing the factory system, was largely responsible for the degradation of the pottery art; Wedgwood wares have usually been in good taste even if they have not always been particularly adventurous. A far more malign influence was that of John Rose of Coalport (Salop). Rose admired the work of Sèvres and imitated it, buying or borrowing specimens to copy and using such ground colours as the *rose Pompadour.* He was one of the first English exponents of the revived Rococo style, which appeared about 1825, and made much porcelain encrusted with applied flowers. His work has been erroneously regarded as a close copy of old Sèvres. Coalport flower painting, however, is very fine in quality and much in the style of Billingsley, who actually worked at the factory for some years.

Spode and Minton

Josiah Spode II, who with his father invented the standard English bone china about 1800, at first made good use of it. Some of his later wares, however, became increasingly pretentious copies of French styles, with highly coloured grounds, lavish gilding, and an excess of applied ornament. In about 1813 William T. Copeland became a partner in the firm, and in 1847 his son, William T. Copeland, Jr., took sole charge of it. In 1970 the company name became Spode, Ltd.

The firm of Minton's was founded at Stoke-upon-Trent in 1793 by Thomas Minton, a Caughley engraver said to have devised for Spode the Broseley Blue Dragon and Willow patterns that are still in use. Like Coalport, the factory was much occupied in copying the work of Sèvres. From 1848 to 1895 they employed a Frenchman, Joseph-François-Léon Arnoux, as art director, and under his tutelage French artists were brought to England; for example, the sculptor Albert Carrier-Belleuse and also Marc-Louis Solon, who was responsible for introducing *pâte-sur-pâte* decoration into England (see below *The European continent*).

The Derby tradition of fine painting was carried into the 19th century, during which time the flower designs became somewhat overblown, although landscapes remained on a high level. The sets of so-called Campaña vases (more properly Campagna), distantly derived from Italianate copies of the Greek krater, were often decorated with landscapes by the brothers Robert and John Brewer and others. The Brewers were pupils of the topographical painter Paul Sandby.

About 1840 Parian ware, an imitation of Sèvres biscuit porcelain, was introduced by Copeland & Garrett (formerly Spode), and a great many figures, some of them extremely large, were made in this medium. Most of them are either sentimental subjects or quasierotic nudes, which were popular subjects of Victorian art. Parian ware had

some success in America, where it was manufactured by Norton and Fenton.

Stoneware and earthenware. Production of earthenware and stoneware for the cheaper market continued on an ever-increasing scale. Lustre decoration, which had been revived in the preceding century, was used more frequently than before. A type of stoneware obviously inspired by Wedgwood's jasperware was made at Castleford, Yorkshire. Ironstone china, a type of opaque stoneware sometimes called opaque porcelain, was introduced early in the 19th century. Pseudo-Chinese and Japan patterns were frequently used to decorate it.

Ironstone china

By 1830 new underglaze colours had been pressed into service for transfer printing. These new colours were particularly used by Ridgway & Co. of Hanley, Staffordshire. Transfer-printed earthenware in blue, which became increasingly popular after 1810, was soon being produced in enormous quantities. It was much used by Spode, who often employed American subjects for wares exported to the United States. Polychrome transfer printing, essayed tentatively at Liverpool during the 1760s, was also mastered.

Earthenware figures were made in large quantities in Staffordshire and elsewhere, the best associated with Enoch Wood. They were intended as chimney ornaments, and the subjects range from bullbaiting to sentimental shepherdesses. Many of them are copied more or less directly from Derby porcelain figures, and they are a sad but accurate reflection of the times during which they were made.

The Great Exhibition of 1851 completed the degeneration started by the revival of the Rococo style. Technical progress allowed the manufacturers ever-increasing elaborations with which they bludgeoned the few remaining sensibilities of their customers. Past styles were indiscriminately and ignorantly copied. Minton's, for example, made an earthenware decorated with coloured glazes that they miscalled maiolica. It was used not only for decorative wares but for domestic articles—such as umbrella stands—and for architectural purposes.

The Paris exhibitions of 1867 and 1878 brought Japanese pottery and porcelain once more to the attention of European manufacturers, but it was not the superb porcelain of Arita that had had so much influence in the previous century. This time the Japanese exported cream-coloured earthenware with a closely crackled surface and lavish painting of poor quality, judging that it would appeal to Western taste. It became extremely popular under the name of Satsuma and was copied avidly at Worcester and elsewhere (see below *Japan: 19th and 20th centuries*).

By 1860 a few people had become profoundly disturbed by the level to which popular taste had sunk. Among them was the English poet and designer William Morris, who founded a firm of interior decorators and manufacturers in 1861. One of his pupils, William de Morgan, started a pottery at Fulham (London) in 1888 that made dishes and tiles inspired by Persian, Hispano-Moresque, and Italian wares. De Morgan used brilliant blues and greens and a coppery red lustre. His designs are a great improvement on those of the factories, although they, too, are derivative.

After about 1860 Doultons of Lambeth (London) copied 18th-century brown stoneware, making small figures and repeating earlier designs. The incised decoration by Hannah Barlow is both pleasant and competent. From a Fulham pottery owned by the Martin brothers came grotesque and often amusing stoneware vases that were sometimes decorated with coloured slips.

Meissen and Sèvres in the 19th century

The European continent. In the 19th century Meissen and Sèvres continued to be the two principal factories and leaders of fashion, although at both places, as elsewhere, artistic standards declined considerably.

In the first half of the 19th century Meissen adopted the revived Rococo style, and a large export trade with England was renewed. This was the period of the sentimental Dresden shepherdess, formerly much admired in England and the United States. Later productions include large and ornate candelabra, overdecorated mirror frames, clock cases, and the like, as well as vases and tureens based on the old Rococo models.

From about 1870, styles altered somewhat and are afterward referred to as those of *die Neuzeit* ("the New Period"). Some of the figures and groups illustrating contemporary subjects throw an amusing sidelight on manners and customs of the time.

At Sèvres, as a result of Napoleon's campaign in Egypt and the newly aroused interest in that country, the Empire style of the first decades of the 19th century incorporated many Egyptian motifs, which were somewhat incongruously translated into porcelain. Also produced were many porcelain plaques with minutely detailed overglaze painting in imitation of easel pictures.

Technical improvements include the introduction, about 1855, of *pâte-sur-pâte,* a process later popular in England, particularly at Minton's. The design was painted in white slip onto a surface of coloured, lightly fired clay. After each coat of slip dried, another was superimposed upon it, until the desired degree of relief had been attained. Finally, it was scraped, smoothed, and incised by metal tools, and the whole object was glazed and fired.

The sculptor Auguste Rodin was employed at Sèvres for a short time but does not seem to have left any enduring marks of his presence. Artistically speaking, Sèvres porcelain has not been very distinguished since the 18th century.

Royal Copenhagen

The Royal Porcelain works at Copenhagen has made a great deal of porcelain with simple patterns in underglaze blue derived from Chinese sources by way of Meissen. Molded fluted shapes are characteristic. Production of the well-known biscuit figures after the sculptor Bertel Thorvaldsen (1768–1844) began in 1867. The factory later introduced a slightly amber-coloured biscuit that was used for figure modelling. Painting on a grayish-toned crackled glaze led to experiments with celadons since, technically, the two have much in common. Other glazes inspired by early Chinese work followed. The firm of Bing and Gröndahl was established in 1853 and has done excellent and imaginative work.

The United States. Although Andrew Duché had succeeded in making porcelain as early as 1741, the first man to produce porcelain in any quantity was William Ellis Tucker of Philadelphia. At first he was a decorator of whiteware, but he started to manufacture both creamware and bone china about 1826. Judge Joseph Hemphill became a partner in 1832, and workmen were imported from Europe. Copies of Sèvres porcelain and other European wares were made about this time in a fine white porcelain body. The first factory at Bennington, Vermont, founded by Capt. John Norton in 1793, made domestic wares, including salt-glazed stoneware. The factory was removed to Bennington Village by his son, Judge Luman Norton, in 1831, and creamware and a brown-glazed ware were produced. In 1839 the factory became Norton and Fenton, and about 1845 the manufacture of Parian ware began. This unglazed near-white porcelain named after Parian marble had been made first in England by Copeland & Garrett (see above *Britain*). John Harrison of Copeland's was hired by Norton and Fenton and brought with him a number of molds. An ironstone china called graniteware or white granite was also made.

Rockingham and yellow-glazed ware

The East Liverpool, Ohio, industry was established in 1838 by James Bennett, an English potter. The first products made there were Rockingham and yellow-glazed ware. In the decade following the Civil War, William Bloor, Isaac W. Knowles, and others introduced the production of whiteware. By the last decade of the 19th century, production had grown until it was the largest pottery-producing area in the world.

At about the same time, Zanesville, Ohio, was also developing as a pottery centre. First production was salt-glazed and slip-decorated stoneware. At a later date much artware was produced in Zanesville plants operated by Samuel Weller, J.B. Owens, George Young, and others. This artware established the basis for a sizable modern interest in collecting. Another important centre during the 19th century was at Trenton, New Jersey, where the first factory was established in 1852. Connected with it was William Bloor, who had some responsibility for putting the industry on a successful footing in East Liverpool. Trenton, like East Liverpool, produced fine, skillfully decorated whiteware.

A close study of the technical side of manufacture was

Study of pottery manufacture at universities

not undertaken until Edward Orton, Jr., succeeded in getting support for the establishment of a department of ceramics at Ohio State University in Columbus in 1894. The New York State College of Ceramics at Alfred, New York, was started soon afterward, with Charles F. Binns as its director. Binns was a member of an English family connected with the manufacture of porcelain at Worcester and Derby during the 19th century and had himself held a supervisory position at Worcester. Similar departments were added to other universities soon afterward, and in 1898 Orton took the lead in forming the American Ceramic Society. In this way knowledge was put on a more scientific basis, and the trained potters who soon became available to the industry were responsible for many technical improvements. Nevertheless, the artistic direction of the factories did not reach a high standard.

Toward the end of the century it became fashionable for American women to study the art of painting on European pottery, and the Cincinnati Art Pottery Company was founded in 1879 to promote sound pottery design. As a result of its work, the Rookwood Pottery was established in 1880 by Maria Longworth Storer. Rookwood wares show a distinct Japanese influence and have excellent red and yellowish-brown glazes (Figure 132).

20TH CENTURY

Pottery factories. At the beginning of the 20th century the Wedgwood factory, whose work has always remained at a high level, extended its already considerable business in the United States, and a service of nearly 1,300 pieces was supplied to the White House during the presidency of Theodore Roosevelt (1901–09). In 1940 the factory began to move to its present site at Barlaston, Staffordshire, after which the historic site at Etruria, Staffordshire, was progressively abandoned.

Designs of Dorothy Doughty

The designs of Dorothy Doughty for the Worcester Royal Porcelain Company, in England, and those of Edward Marshall Boehm, at Trenton, New Jersey, established a new development in decorative porcelain. Characteristic of this kind of work are the American birds of Dorothy Doughty issued in limited editions by the Worcester Company. They are especially remarkable for technical advances in preparing the article for firing, which allow the material to be treated with much greater freedom than hitherto. Porcelain becomes very soft when it reaches the point of vitrification, but, using an elaborate series of props to support free-floating parts, the Worcester technicians succeeded in firing designs that would have been completely impossible earlier. Associated with these models are exact reproductions of natural flowers that also excel in complexity and verisimilitude anything made in the past.

In the early part of the 20th century, Bernard Moore experimented with Chinese glazes (see below *China: Ch'ing dynasty*). He produced some successful flambé and *sang-de-boeuf* glazes on a stoneware body at his small factory in Stoke-upon-Trent. He worked in association with William Burton of Pilkington pottery in Manchester, which made experimental decorative ware of all kinds.

After World War I, figure modelling worthy of the old Meissen tradition was done by Paul Scheurich, Max Esser, Paul Börner, and others. The early red stoneware was also revived. This renaissance was halted temporarily by World War II, but production was resumed by 1950. The wares exported from the German Democratic Republic into western Europe were excellent in quality.

A factory that has preserved its traditional reputation for fine porcelain is Nymphenburg, at Munich, now the Staatliche Porzellan-Manufaktur Nymphenburg. At the beginning of the 20th century it began to use a wider range of underglaze colours with the aid of colour chemists from Sèvres and, about the same time, reissued some of the old figures and services of Bustelli and Auliczek (appropriately marked). Attention was soon turned to services of fine quality in the modern idiom, and excellent figures by Resl Lechner and others were produced. Lechner succeeded in adapting the 18th-century styles to 20th-century purposes in a manner that is an object lesson to those manufacturers who insist, even today, on adding the scrolls and flourishes of the Rococo.



Figure 132: Vase painted in coloured slips under glaze, Rookwood pottery, Cincinnati, Ohio, c. 1900. In the Victoria and Albert Museum, London. Height 26.4 cm.

By courtesy of the Victoria and Albert Museum, London; photograph, A.C. Cooper Ltd

Such factories as Rörstrand and Gustavsberg in Sweden and Arabia Oy in Finland have achieved a growing reputation for excellent design in the modern idiom (Figure 133). The emphasis on form in present-day pottery is to a great extent due to the import of Chinese wares of the Sung dynasty (see below) during the 1920s.

The pottery of the United States bears comparison with that of any other country, and standards are constantly improving. Technically, the United States is perhaps ahead of much of the rest of the world. The growing appreciation of good pottery design has led the national government, as well as state and local governments, to sponsor pottery making as an art. There is also a pottery experimental station in the Tennessee Valley.

Government-sponsored pottery making

The artist-potter. The artist-potter has had an important influence on modern design from the time that Bernard Leach (1887–1979) established the Leach Pottery in St. Ives, Cornwall, in 1920. Leach spent many of his early years in the Far East and learned the art of making *raku* and stoneware in Japan (see below *Japan: Azuchi-Momoyama period*). He began working at a time when

By courtesy of the Victoria and Albert Museum, London; photograph, A.C. Cooper Ltd



Figure 133: "Grey Bands," porcelain service designed by Wilhelm Kåge for the Gustavsberg factory, Sweden, 1944. In the Victoria and Albert Museum, London. Height of jug 14 cm.

interest in early Chinese wares had greatly increased, and much of his work is obviously influenced by the work of Tz'u-chou (see below *China: Sung dynasty*), as well as that of Japan. It is, nevertheless, strongly individual. One of Leach's pupils, Michael Cardew, has done good work in stoneware, which he often decorated with vigorous patterns drawn with a pleasing economy of outline (Figure 134). William Staite Murray, at one time the head of

By courtesy of the Victoria and Albert Museum, London; photograph, A.C. Cooper Ltd



Figure 134: Slipware jug with clear, honey-coloured glaze by Michael Cardew, Winchcombe, Gloucestershire, *c.* 1938. In the Victoria and Albert Museum, London. Height 29.5 cm.

the ceramic department of the Royal College of Art, has made some important and interesting stoneware and has influenced many younger potters. Excellent work has been done by continental potters working in England, among them Lucie Rie from Vienna and Hans Coper from Germany. Amusing figures have come from Marion Morris, who was trained in Budapest.

The artist-potters on the Continent tend to be less conservative than their English counterparts, and many new and interesting developments have occurred. Abstraction is particularly favourable to development, since the potter understood its principles long before the 20th-century painters and sculptors came to it. By the 1970s many art schools included pottery making in their curriculum, and students were increasing in numbers.

Somewhat outside the mainstream of pottery tradition, and a markedly individual production related to their work in other media, is the pottery of such well-known artists as Pierre-Auguste Renoir, Paul Gauguin, Joan Miró, Henri Matisse, Ernst Barlach, and Pablo Picasso. A good many of these wares are unique, although some have been repeated by factory production methods.

East Asian and Southeast Asian pottery

CHINA

Nowhere in the world has pottery assumed such importance as in China, and the influence of Chinese porcelain on later European pottery has been profound.

It is difficult to give much practical assistance on the question of Chinese marks. Most of the Chinese marks give the name of the dynasty and that of the emperor; however, many of them have been used so inconsequentially that, unless the period can also be assigned with reasonable certainty by other means, it is better to disregard them. The dating of Chinese pottery is further complicated by the fact that there were traditional and persisting types that overlapped; quite often, therefore, dynastic labels cannot be regarded as anything more than an indication of the affinities of the particular object under discussion.

Chinese symbolic decoration

Chinese decoration is usually symbolic and often exploits the double meaning of certain words; for instance, the Chinese word for bat, *fu,* also means "happiness." Five bats represent the Five Blessings—longevity, wealth, serenity, virtue, and an easy death. Longevity is symbolized by such things as the stork, the pine, and the tortoise, the *ling chih* fungus, and the bamboo, all reputed to enjoy long life. The character *shou,* which also denotes longevity, is used in a variety of ornamental forms. Together, the peach and the bat represent *fu-shou,* long life and happiness. The "Buddha's hand" citron, a fruit with fingerlike appendages, is a symbol of wealth, and each month and season is represented by a flower or plant. The *pa kua,* consisting of eight sets of three lines, broken and unbroken in different combinations, represent natural forces. They are often seen in conjunction with the Yin-Yang symbol, which represents the female-male principle, and which has been well described by the pottery scholar R.L. Hobson as resembling "two tadpoles interlocked." The dragon generally is a mild and beneficent creature. It is a symbol of the emperor, just as the phoenix-like creature (*feng-huang*) symbolizes the empress.

There are three principal religious systems in China: Confucianism, Taoism, and Buddhism. Taoist figures, in particular, appear frequently on porcelain as decoration. The most important, Lao-tzu, has a large and protuberant forehead. He is usually accompanied by the "eight immortals" (*pa hsien*), and these are sometimes modelled as sets of figures. The eight horses of the emperor Mu Wang (Chou dynasty) are also frequently represented. The Buddhist goddess Kuan-yin and the 18 lohan, disciples of Buddha, were also modelled. The "eight Buddhist emblems" appear fairly frequently, as do the "eight precious things" and a collection of instruments and implements used in the arts and known as the "hundred antiques." The "lions of Buddha" (often miscalled dogs) are frequently represented, as is the kylin (properly *ch'i-lin*), which is a composite animal, not unlike a unicorn, that has a fierce appearance but gentle disposition.

Most of these symbols were not used in pottery decoration before the Ming dynasty, although both the dragon and a phoenix-like creature (probably the Chinese pheasant), as well as some floral motifs, are earlier. The *lei-wen,* however, which resembles the Greek key fret (an ornament consisting of small, straight bars intersecting one another in right angles) and is sometimes used on the later ceramic wares, appears on bronzes as early as the Shang and Chou dynasties, where it is called the cloud-and-thunder fret. The *t'ao-t'ieh,* which is a grotesque mask of uncertain origin, also appears on early bronzes and on later pottery and porcelain. Decorations based on Chinese literary sources are usually extremely difficult to trace to their origin.

Earliest Chinese pottery

The earliest Chinese pottery is of the Neolithic period and has been discovered in the provinces of Honan and Kansu. Perhaps the best known of these wares is a series of large urns of red polished pottery with geometric decoration found in the Pan-shan cemetery and at Mach'ang, both in Kansu Province. These were made by hand, the latest specimens with perhaps some assistance from a slow wheel, and are at least as early as 2000 BC.

The only known complete specimen of a fine white stoneware dating from *c.* 1400 BC (Freer Gallery of Art, Washington, D.C.) is decorated with chevrons (linked V-shapes) and a key-fret pattern, the shoulder motifs being reminiscent of those seen on contemporary bronze vessels. This ware is much better in quality than most other surviving pottery of the Shang period (18th to 12th century BC) or of the following Chou dynasty (1111–255 BC). Much Chou pottery is decorated with rudimentary incised ornament, some of which resembles the impress of coarse textiles referred to as mat markings. The shapes used for these pieces were often inspired by bronze vessels.

The development of glazing in China may have started with the application of glass paste to some of the later Chou wares. Stoneware vessels of about the 3rd century BC have a glaze that is little more than a smear but one that has obviously been deliberately applied. This type persisted for several centuries.

The first pottery to survive in appreciable quantities belongs to the Han dynasty (206 BC–AD 220); most of it has been excavated from graves. Perhaps the commonest form is the *hu*, a baluster-shaped vase copied from bronze vessels of the same name and sometimes decorated with relief ornament in friezes taken directly from a bronze original. The hill jar, which has a cover molded to represent the Taoist "Isles of the Blest," is another fairly frequent form, and many models of servants, domestic animals, buildings, wellheads, dovecots, and the like also have been discovered in graves. Some of this pottery is unglazed or decorated with cold (*i.e.*, unfired) pigments, but much of it is covered with a glaze that varies from copper green to yellowish brown; often the colours have become iridescent from long burial. The body is usually a dark red and approaches stoneware in hardness.

Han glaze is more glasslike than that of the Chou period and is of an excellent quality. It contains lead and was frequently coloured green with copper oxide.

Celadon glazes

Yüeh *yao* ("Yüeh ware") was first made at Yüeh-chou, Chekiang Province, during the Han dynasty, although all surviving specimens are later, most belonging to the Six Dynasties (AD 222–589). They have a stoneware body and an olive or brownish-green glaze and belong to the family of celadons, a term that looms large in any discussion of early Chinese wares. It is applied to glazes ranging from the olive of Yüeh to the deep green of later varieties. These colours were the result of a wash of slip containing a high proportion of iron that was put over the body before glazing. The iron interacted with the glaze during firing and coloured it.

T'ang dynasty (AD 618–907). Chinese pottery reaches an important stage in its development during the T'ang dynasty.

Nearly everything that has survived has been excavated from tombs, many of them found accidentally by railway engineers and latterly by more systematic excavations. Excavations at Sāmarrā' on the Tigris, a luxurious residence built by the caliph al-Mu'taṣim (son of Hārūn ar-Rashīd) in AD 836 and abandoned in 873, have uncovered many fragments of T'ang wares of all kinds. Perhaps the most important finds from a historical viewpoint are the fragments of what is undoubtedly porcelain. An Islāmic record of travels in the Far East, written in 851, records "vessels of clay as transparent as glass." There can be little doubt, therefore, that translucent porcelain was made in the T'ang period, although it was not until the Yüan dynasty (1206–1368) that it began to resemble the type with which the West is most familiar.

Early porcelain

Perhaps the most important single development was the use of coloured glazes—as monochromes or splashed and dappled. The T'ang wares commonest in Western collections are those with either monochrome or dappled glazes covering a highly absorbent, buff, earthenware body. The dappled glazes were usually applied with a sponge, and they include blue, dark blue, green, yellow, orange, straw, and brown colours. These glazes normally exhibit a fine crackle and often fall short of the base in an uneven wavy line, the unglazed surface area varying from about one-third to two-thirds of the vessel.

Dappled glazes are also found on the magnificent series of tomb figures with which this period is particularly associated (Figure 135). Similar figures were made in unglazed earthenware and were sometimes decorated with cold pigment. Although the unglazed specimen or those covered only with the straw-coloured glaze are occasionally modelled superbly, many are crude and apparently made for the tombs of the less affluent and influential. Most of the glazed figures are much better in quality and occasionally reach a large size; figures of the Bactrian camel, for instance, are particularly impressive, some being nearly three feet high. The Bactrian pony, introduced into China about 138 BC, is to be found in many spirited poses. This fashion for tomb figures fell into disuse at the beginning of the Sung dynasty (AD 960–1279) but was revived for a short while during the Ming period (1368–1644), when T'ang influence is noticeable.

Marbled wares are seen occasionally. The effect was achieved either by combing slips of contrasting colours



Figure 135: Ceramic tomb figure decorated in characteristic coloured glazes, T'ang dynasty (AD 618–907). In the Victoria and Albert Museum, London. Height 71 cm.
By courtesy of the Victorian and Albert Museum, London

(*i.e.*, mingling the slips after they had been put on the pot, by means of a comb) or by mingling differently coloured clays. Another type of T'ang ware (probably from Honan) had a stoneware body with a dark-brown glaze streaked by pale blue. Most vessels stand on a flat base; although later T'ang wares sometimes were given a foot ring, for the most part this can be regarded as evidence in favour of a Sung dating.

Sung dynasty (AD 960–1279). The wares of the Sung dynasty are particularly noted for brilliant feldspathic glazes over a stoneware body and their emphasis on simplicity of form. Decoration is infrequent but may be incised, molded, impressed, or carved; a certain amount of painted decoration was done at Tz'u-chou in Chihli

By courtesy of the Percival David Foundation of Chinese Art, London



Figure 136: Ting ware vase, white porcelain with an ivory-white glaze covering carved design of peony flowers and foliage, Sung dynasty (AD 960–1279). In the Percival David Foundation of Chinese Art, London. Height 36.8 cm.

(now called Hopeh) Province (see below). The esteem accorded to the Sung wares accounts for the relatively large number that have survived. The principal varieties are Ju, *kuan*, Ko, Ting, Lung-ch'üan, Chün, Chien, Tz'u-chou, and *ying-ch'ing*.

Ju ware has a buff stoneware body and is covered with a dense greenish-blue glaze that sometimes has a fine crackle. It was made in Honan at an Imperial factory that apparently had a life of about 20 years, starting in 1107.

Kuan ("official") is another Imperial ware that is also exceedingly scarce. It was probably first made in the north, the kilns being reestablished at Hangchow in Chekiang Province about 1127, when the court fled southward to escape the Chin Tatar invaders. The body is of stoneware washed with brown slip. The glaze varies from pale green to lavender blue, with a wide-meshed crackle emphasized by the application of brown pigment. Chinese references to "a brown mouth and an iron foot" can be identified with the colour of the rim and the foot ring.

Ko ware is closely related to *kuan* ware. It has a dark stoneware body and a grayish-white glaze with a well-marked crackle, which was induced deliberately for its decorative effect.

Ting wares are white. Some exhibit an orange translucency (Figure 136), while the coarser varieties are opaque. The finest examples are called "white" (*pai*) Ting. On the exterior of bowls and similar vessels the glaze of white Ting is apt to collect in drops, called teardrops. Many articles, particularly bowls, were fired mouth downward, leaving an unglazed rim that was afterward bound with a band of copper or silver. (Bands appear occasionally on other Sung wares, notably *ying-ch'ing*, and were sometimes used to conceal damage rather than an unglazed rim.) Coarser varieties are known as "flour" (*fen*) Ting and "earthen" (*t'u*) Ting, and there are also a few examples of black Ting. As in the case of *kuan* ware, the kilns are said to have been removed southward in 1127, but it has so far proved impossible to differentiate between the northern and southern varieties. Other white wares made elsewhere during the period include those of Tz'u-chou and a variety covered with a white slip over a grayish body from Chü-lu Hsien (both in present Hopeh Province).

Lung-ch'üan

The celadons of Lung-ch'üan are, perhaps, the most common of the classic Sung wares (Figure 137). The town is in the province of Chekiang, near the capital of the southern Sung emperors at Hangchow. The kilns probably date back to the 10th century. The glaze, of superb quality, is a transparent green in colour. It is thick and viscous, usually with a well-marked crackle. (The glaze on early specimens is less transparent and is denser.) The body is gray to grayish white, best seen at the rim, where the glaze tends to be thin. By far the most frequent surviving examples of Lung-ch'üan celadon are large dishes, for which there was a thriving export trade, due in part to the superstition that a celadon dish would break or change colour if poisoned food were put into it. Bowls and large vases, both of which are scarce, were also made with this glaze. Decoration is usually incised, but molded decoration is also found. On some pots the molding was left unglazed, so that it burned to a dark reddish brown—an effective contrast to the colour of the glaze. The more finely potted wares are the scarcest and often the oldest. The heavier varieties were intended to withstand the rigours of transport to overseas markets, and probably most of them belong to the Yüan dynasty (1206–1368), when the export trade was considerably extended.

By courtesy of the Victoria and Albert Museum, London



Figure 137: Lung-ch'üan celadon ware wine jar and cover, with light bluish-green glaze, Sung dynasty, 12th century. In the Victoria and Albert Museum, London. Height 25.4 cm.

Chün ware comes from the K'ai-feng district of Honan Province. The body is a grayish-white, hard-fired stoneware covered with a thick, dense, lavender-blue glaze often suffused with crimson purple. This is the first example of a reduced copper, or flambé, glaze. Conical bowls are especially numerous, and dishes are not unusual, but the finer specimens are usually flowerpots, sometimes said to have been made for Imperial use. Characteristic are barely perceptible channels or tracks caused by the parting of the viscous glaze; these are called earthworm tracks by the Chinese. The kilns probably continued to produce this ware until the 16th century, and it is difficult to separate some of the later productions from the earlier.

Chien ware is called after the original place of manufacture, Chien-an, in Fukien Province. Manufacture was later moved to nearby Chien-yang, probably during the Yüan period. The glaze is very dark brown, approaching black, over a dark stoneware body, and it usually stops short of the base in a thick treacly roll.

There are many variations in the colour of the glaze. Streaks in lighter brown are referred to by the Chinese as hare's fur. Silvery spots on the glaze are called oil spots. The most usual surviving form is the teabowl; these were much esteemed by the Japanese under the name of *tenmoku* and were used in the tea ceremony (see below *Japan: Kamakura and Muromachi periods*).

Tz'u-chou ware

The kilns of Tz'u-chou, formerly in Honan, are now in Hopeh Province. The earliest surviving examples are referable to the T'ang dynasty. In the Sung period, vases, wine jars, and pillows (which are more comfortable than they appear) were the most usual products. The body is usually a hard-fired, grayish-white stoneware that was first covered with a wash of white slip and then with a transparent glaze. For the first time painted decoration appears under the glaze, perhaps as a result of influence from the Middle East (see above *Islāmic: Mesopotamia and Persia: 11th to 15th century*). Decoration is nearly always in brown or black; the motifs are usually floral and display a singular freedom of line that is very attractive. (The inclusion of human and animal figures suggests a Yüan or a Ming dating, at the least.) The slip covering was sometimes carved away, leaving a pattern in contrasting colour, a technique also used in conjunction with a dark brown glaze. A hare's-fur glaze, similar to that of Chien wares, was also employed. A blue glaze with painted decoration in black beneath it was obviously inspired by contemporary Persian pottery decorated in the same way. Another innovation, perhaps derived from the same source, is the use of colours applied over the glaze. These are limited to primitive reds and greens and yellows.

An important and not uncommon ware is *ying-ch'ing* ("shadowy blue"). It was manufactured in both the south (Kiangsi) and the north (Hopeh). Moreover, it was extensively exported and has been found as far west as the ruins of al-Fusṭāṭ in Old Cairo. The body is pale buff in colour, usually translucent, and thinly potted, breaking with a sugary fracture. Most genuine examples seem to belong to the Sung and Yüan periods, but it is probable that, in the north at any rate, manufacture started late in the T'ang

dynasty and lasted well into the Ming period. Bowls of conical form are the commonest survival, and many are decorated with incised floral and foliate motifs. Lightly molded decoration occurs, as does combing of the clay. The *mei p'ing* vase is found with this glaze; it has a tall body with straight sides, high, rounded shoulders, and a short narrow neck and was intended to hold a single spray of prunus blossom. Stem cups, deep bowls, and ewers were also produced. Bowls sometimes have the rim bound with copper.

Yüan dynasty (1206–1368). The Yüan, or Mongol, dynasty is often regarded as being no more than transitional between Sung and Ming types. This is not entirely true. Undoubtedly, many Sung types were continued, just as the T'ang types were continued at the beginning of the Sung dynasty, but there are other wares that represent a new departure. The manufacturing centre of Ching-te-chen increased in importance and first manufactured the white translucent porcelain that was to have a revolutionary effect on Chinese wares. The use of painted decoration, begun during the Sung period at Tz'u-chou, also became much more widespread, and the two techniques were combined in a manner that later affected the course of porcelain manufacture throughout the world.

The *Ko ku yao lun* of 1387 refers to *shu fu* ware, a type of white porcelain. The base is unglazed. Decoration in relief, painted in slip or engraved, is to be seen on some surviving examples of porcelain. It appears to be a development of the earlier *Ting ch'ing*. Much more unusual is the appearance of a few specimens of Yüan date that are painted with reduced copper red under the glaze. As mentioned above, the potters of Chün-chou had achieved this colour, but only in the glaze.

The use of underglaze blue was introduced from the Middle East, where it had been employed at least as early as the 9th century, specimens thus decorated having been recovered at Sāmarrā' (see below *Islāmic pottery: 'Abbāsid*). The best known example of Yüan porcelain decorated in this manner, which is usually referred to as blue-and-white, is a pair of vases in the Percival David Foundation of Chinese Art in London. They bear a date equivalent to 1351. The peony scroll, carved or in applied relief, appears on some of these blue-and-white wares.

Blue-and-white ware

Ming dynasty (1368–1644). The Mongol emperor Shun Ti was defeated in a popular uprising, and the Hungwu emperor, founder of the Ming dynasty, succeeded him in 1368. When the country had recovered from these internecine struggles, pottery art took a new lease of life, though under somewhat changed conditions. The Sung wares went out of favour, and the old factories sank into obscurity, while the fame and importance of the great porcelain town of Ching-te-chen, near the P'o-yang Hu in Kiangsi Province, overshadowed all the rest. The Imperial factory there was rebuilt and reorganized to keep the court supplied with the new porcelain. The staple product of Ching-te-chen was the fine white porcelain that made "china" a household word throughout the world; and as this ware lent itself peculiarly well to painted decoration, the vogue for painted porcelain rapidly replaced the old Sung taste for monochromes.

"China" ware

The reign of the Yung-lo emperor (1402–24) is remarkable for some extremely thin-walled pieces, referred to as "bodiless" (*t'o t'ai*) ware. Engraved examples are known, and Chinese commentaries refer to specimens decorated in red.

After this early period, Ming wares generally are fairly easily recognizable. Porcelain replaced stoneware as the usual medium, and polychrome decoration became widely employed. The largest single group of Ming porcelain is that painted in blue underglaze (Figure 138). Much of the pigment used was imported from Middle Eastern sources. Supplies of this so-called Mohammedan blue (*hui-hui ch'ing*), which came from the Kashān district of Persia, were not always obtainable and were interrupted on more than one occasion. The quality of the blue-painted wares, however, remained to a great extent dependent on its use until the end of the 16th century, when methods of refining native cobalt were devised.

The wares lack much of the precision of the porcelain made during the following Ch'ing period (1644–1911/12), when a kind of factory system grew up that divided the work into a large number of repetitive operations. Little trouble was taken to smooth over imperfections of manufacture, and foot rings are often finished summarily. The glaze, too, frequently has minor defects, and articles, such as vases, are sometimes slightly distorted and carelessly finished. The shape of many examples can fairly be described as massive, in spite of the fact that most of them were made for export, and the difficulties of transporting them must have been considerable. None of these factors evinces a lack of skill, especially as the potters were quite capable of technical virtuosity when they wished to display it—some of the most thinly potted of all Chinese porcelain belongs to this period. It seems that the Ming potters disdained the attitude of mind that treated blemishes as important; occasional distortions, in fact, were regarded as lending interest to an object. The Chinese did not carry this aesthetic creed to the same lengths as the Japanese (see below *Japan*), but the difference seems to be largely one of degree. Ming wares can fairly be described as masculine, in contrast to the more feminine, more precisely finished wares of the later Ch'ing period.

By courtesy of the Victoria and Albert Museum, London



Figure 138: Flask decorated with a dragon and wave scrolls in underglaze blue, Ming dynasty, 14th century. In the Victoria and Albert Museum, London. Height 36.8 cm.

Reign of the Hsüan-te emperor (1425–35). In this period the arts were particularly fostered and a high level of achievement attained. The blue painting is blackish in colour, with dark spots at intervals where thick blobs of pigment were deposited by the brush—the so-called heaped and piled effect. The motifs are floral and foliate, with the occasional use of fish and waterfowl. Sometimes vessels are bordered by a pattern of conventional rock amid waves—the Isles of Immortality—often referred to as the Rock of Ages pattern. The pattern appears frequently throughout the Ming period and later.

Heaped and piled effect

Contemporary Chinese commentaries refer to the use of underglaze copper red—often called sacrificial red for uncertain reasons. To a great extent sacrificial red was abandoned later in the dynasty in favour of overglaze iron red, although it was used again during the reign of the Ch'ing dynasty K'ang-hsi (1661–1722) and Yung-cheng (1722–35) emperors and appears in a rather primitive form from some provincial kilns. Both copper red and blue were used as monochromes and, occasionally, together; but since these pigments required a slightly different firing temperature, one or the other is usually deficient in quality.

The use of overglaze colours was rare, and the technique had by no means been fully mastered.

Reign of the Ch'eng-hua emperor (1464–87). Much

overglaze decoration can be attributed with a reasonable measure of certainty to the reign of Ch'eng-hua, the finest examples being, perhaps, the chicken cups, so-called because they are decorated with chickens. Their decoration is outlined in underglaze blue and filled in with soft overglaze colours called "contending colours" (*tou ts'ai*). Ch'eng-hua overglaze colours were thin, subdued in colour, and pictorial in effect.

The practice of enamelling directly onto unglazed, or biscuit, porcelain instead of onto a glazed and fired body is sometimes thought to have begun in this reign, though that of the Chia-ching emperor (1521–67) is the more likely. Ming specimens are, in any case, extremely rare; most belong to the reign of the K'ang-hsi emperor (1661–1722) in the Ch'ing dynasty.

Reigns of the Hung-chih and Cheng-te emperors (1487–1521). The first use of a coloured overglaze ground can be attributed to the reign of the Hung-chih emperor (1487–1505), when a yellow of variable shade first appears. In the reign of the Cheng-te emperor (1505–21) the influence of the Muslim palace eunuchs who supervised the Imperial kilns is seen in such blue-and-white motifs as the Mohammedan scroll, which is composed of somewhat formal flowers joined by S-shaped stems, with scroll-like leaves at intervals along them. Mohammedan blue was again available. The earliest versions of this theme, which seems orginally to have come from a textile pattern, are the least stiffly drawn. The linear style of painting characteristic of earlier porcelain altered to one in which outlines were filled in with flat ungraduated washes.

Muslim influence

Reign of the Chia-ching emperor (1521–67). This reign is notable for a deterioration in the quality of the porcelain body, offset by the use of rich dark blue. Wares painted overglaze, too, were executed in good colours, with well-marked outlines. A characteristic colour, the opaque iron red (*fan hung*), sometimes called tomato red, was used as a monochrome with gilt traceries over it on bowls that sometimes had interior decoration in underglaze blue. Various wares have decoration in red and green, a palette that became more familiar later. A yellow glaze is found in conjunction with incised decoration (usually a dragon) in green. Very rarely was a green or blue monochrome used.

Reigns of the Lung-ch'ing and Wan-li emperors (1567–1620). The styles of Chia-ching were, to some extent, continued in the following reigns of the Lung-ch'ing (1567–72) and Wan li (1572–1620) emperors. A palette containing underglaze blue in conjunction with green, yellow, aubergine purple, and iron red (the precursor of the later Ch'ing *famille verte* palette) was known as "Wan-li five-colour" ware (Wan-li *wu ts'ai*). The red and green Chia-ching decoration was also used, and vast quantities of blue-and-white porcelain were produced for export. The body is quite unlike that used earlier in the dynasty, being thin, hard, crisp, and resonant. It is the commonest of all Ming wares in the West. During the reign of the Wan-li emperor, much pierced work (*ling lung*) was done. Pierced objects range from small brush pots to vases with coloured glazes sometimes termed *fa hua*.

The Ming dynasty ended in 1644. The wares of the last three emperors, for the most part, followed styles already established; perhaps an exception can be made for blue-and-white, which shows a number of new departures in both form and decoration. Many of the vases are without a foot-ring and stand on a flat, unglazed base. Forms based on European wares were obviously made for export.

Provincial and export wares. Most of the wares hitherto discussed were made in the Ching-te-chen area; it remains to consider the other wares of the period. The export of celadons went on, not only to the countries west of China but also to Japan, where they were much esteemed. Most celadons attributable to the Ming period have incised under the glaze floral and foliate decoration of a kind that also appears on blue painted wares.

The fine porcelain of Te-hua in Fukien Province was first made, perhaps, in the early part of the dynasty. Most of this porcelain was left undecorated, and received the name in Europe of blanc de chine. The glaze is exceptionally thick and lustrous, and early examples are often slightly ivory in tone. Overglaze painting is infrequent; virtually all early coloured speciments, figures or vessels, have been decorated in Europe, usually in the Netherlands. Figures especially were produced at the Te-hua kilns, with the Buddhist goddess Kuan-yin being a favourite subject.

Blanc de chine porcelain

The stoneware of I-hsing in Kiangsu Province was known in the West as Buccaro, or Boccaro, ware and was copied and imitated at Meissen, at Staffordshire, and in the Netherlands by Ary de Milde and others. Its teapots were much valued in 17th century Europe, where tea was newly introduced. The wares of I-hsing are unglazed, the body varying from red to dark brown. The molding is extremely precise and was often sharpened by grinding on a lapidary's wheel. The body was sometimes polished in the same way.

Most of the Ming stoneware ridge tiles and roof finials were made at kilns near Peking. Many of them are decorated in green, yellow, turquoise, and aubergine-purple glazes, recalling the wares of the T'ang dynasty. A Ming date is exceedingly optimistic for most of them. To this group belong, it is thought, a few large figures that have sometimes been somewhat doubtfully awarded a T'ang date.

The provincial tile kilns also manufactured "three-coloured" (*san ts'ai*) wares, perhaps originally a product of the Tz'u-chou kilns. These were decorated with coloured glazes that were often kept from intermingling by threads of clay (cloisonné technique) or were used in conjunction with the pierced technique (*fa hua*). Others have engraved designs under the glazes. Most existing specimens are large vases, barrel-shaped garden seats, and the like. The best are extremely handsome and imposing, turquoise and dark-blue glazes being particularly effective.

Ch'ing dynasty (1644–1911/12). With the Ch'ing dynasty came the beginning of the immense vogue for porcelain in Europe that was to reach its height during the first half of the 18th century. Many varieties of Ch'ing ware are common in the West. Its wares differ, for the most part, from those of the Ming period in a fairly distinctive manner. Potters had their medium under almost complete control, and their products are much more precisely finished. Their finesse contrasts sharply with the struggles of potters in Europe, where porcelain manufacture did not emerge from the purely empirical stage until the 19th century. Letters written in 1712 and 1722 by a Jesuit missionary who spent some years at Ching-te-chen record that some Ch'ing pieces were handled by as many as 70 men, each contributing a small part to the total effect, and this is one of the reasons why many Ch'ing wares are found to lack the freshness and the spontaneity of Ming decoration.

Difference between Chinese and European pottery making

The Imperial kilns of Ching-te-chen were fortunate in the support they received from the palace during the reigns of the K'ang-hsi (1661–1722), Yung-chen (1722–35), and Ch'ien-lung (1735–96) emperors. The K'ang-hsi emperor, in particular, was a patron of the arts on a considerable scale.

Underglaze blue and red. The blue-painted porcelain of the Ch'ing dynasty has been somewhat neglected in the 20th century. This is probably due to the ridiculously high value placed on it during the latter years of the 19th century, when it was often called Nanking ware. Even the best, which belongs to the reign of the K'ang-hsi emperor, hardly bears comparison with the finer Ming wares, though its influence on European porcelain was far-reaching. Blue-and-white porcelain was exported to Europe in vast quantities, and many of the forms were especially made for export; the condiment ledge on plates and dishes, for instance, which first appeared in the reign of the Wan-li emperor (Ming dynasty), had been added for Western customers (the Chinese used the saucer dish). The blue-and-white of the K'ang-hsi period has an extremely white body, and the blue is exceptionally clear and pure. It is variable in shade, and the design is executed in graduated washes within lightly drawn outlines, a point of difference from Ming wares. Many of the designs of the Ming period were in use, and, of the later patterns, those illustrating literary and historical themes are probably of the highest quality.

Ginger jars decorated with prunus blossom reserved in white on an irregular blue ground, intended to represent

the cracked ice of spring and sometimes described as pulsating, were once valued highly; in the mid-20th century a more realistic attitude was taken toward them.

Underglaze copper red was also used during the 18th century. The stem cups of the Yung-cheng period with three fruit or three fish in silhouette, which imitate those of Hsüan-te, are much better known than the wares they copied. Copper red also appears in conjunction with underglaze blue, and a greenish-toned glaze is common with pieces thus decorated.

Underglaze blue was sometimes used as a monochrome ground colour. It was blown on the surface in powder form before glazing; a bamboo tube, closed with gauze at one end, was employed for the purpose. It is thus called powder blue, or, in Chinese, *ch'ui ch'ing* ("blown blue"), and is distinct from the sponged blue grounds of the Ming dynasty. It was subsequently used at several of the porcelain factories in Europe. Clair du lune (*yüeh pai*, "moon white"), a cobalt glaze of the palest blue shade, was also used.

Use of powder blue

Coloured glazes. Copper red, called oxblood (*sang-de-boeuf*) by the French, appears in monochrome form as Lang *yao*. This glaze was also known to the Chinese as "blown red" (*ch'ui hung*). It was certainly used as a monochrome in early Ming times and possibly even earlier, and is the direct ancestor of the showy *flambé* glazes (*yao pien*) (Figure 139) of the Ch'ien-lung period that are often vividly streaked with unreduced copper blue.

By courtesy of the Victoria and Albert Museum, London photograph, A.C. Cooper Ltd.



Figure 139: Vase with *flambé* glaze (*yao pien*) of reduced copper, Ch'ing dynasty, reign of the Ch'ien-lung emperor, 1736–96. In the Victoria and Albert Museum, London. Height 32.4 cm.

Another variation, no doubt at first accidental, is the glaze known in the West as "peach bloom," a pinkish red mottled with russet spots and tinged with green. The Chinese have various names for it, but perhaps the commonest is "bean red" (*chiang-tou hung*). It is used on a white body. Most objects glazed in this way are small items for the writer's table.

Monochromes of all kinds are a distinct and important section of Ch'ing wares, and many reproductions of Sung monochromes were made. The use of iron as a pigment can be seen in a revival of the celadon glaze. The Ching-te-chen celadons have, generally, a pale-green glaze over white porcelain, the foot-ring being given a wash of brown to simulate the old ware. Meanwhile, celadons of the Lung-ch'üan type were still being made. In addition to the celadon glaze, iron was used to produce colours varying between café au lait and pale yellow and also "deadleaf" brown.

Sometimes panels were reserved in white and painted in overglaze colours. Specimens thus glazed appear in the old Dutch catalogues as Batavian ware, because the wares were imported via the Dutch centre of trade and transshipment at Batavia (modern Djakarta), in Java. They are also related to "mirror black" (*wu chin*), a lustrous colour obtained by the addition of manganese, and sometimes decorated with gilding or even, as in at least one extant specimen, with both gilding and silvering. Imperial yellow, a lead glaze often used over engraved dragons and similar designs, was again employed during the 19th century.

Brilliant turquoise glazes derived from copper have been produced up to the mid-20th century, although later examples seldom have the quality of the earlier ones. The glaze is usually covered with a network of fine crackle, and in some examples there is engraved decoration under the glaze. Related glazes are the copper greens—for example, leaf green and cucumber green, the latter being speckled with a darker colour. Apple green is an overglaze colour used as a ground and applied over a crackled gray glaze. Most greens are relatively late.

Purple, or aubergine, glazes derived from manganese are seen occasionally. Brinjal bowls, decorated with engraved flowers, have an aubergine ground in conjunction with dappled green and yellow glazes. (Brinjal, in fact, means aubergine, or eggplant, which is a favourite food in parts of the East.) Bowls with engraved dragons and a combination of only two of these colours are somewhat better in quality.

Overglaze colours. Overglaze colours were sometimes used as monochromes; for example, iron red or, as it is sometimes called because it varies a little in shade, coral red. The surface is usually glossy but occasionally mat. The rose colour, discussed below, was used both as a monochrome and as a ground colour.

The wares enamelled on biscuit are a much sought after group. They are a development of the Ming *san ts'ai* wares, which were still being made during the Ch'ing period. The effect of painting directly on biscuit was to produce a soft and distinctive colouring that is extremely attractive. The outlines were first painted directly on the unglazed surface in brownish black; some of the colours were then painted within these outlines and others were washed over them; however, red or blue overglaze colours, when they appear, are usually provided with a patch of glaze underneath them. The practice seems hardly to have survived the K'ang-hsi period, except for deliberately made later copies.

Famille verte overglaze

During the reign of the K'ang-hsi emperor the wares decorated overglaze were painted in the *famille verte* palette, usually over a white glaze (Figure 140). The name *famille verte* ("green family") is derived from the distinctive green employed, but the wares are a development of the Wan-li five-colour ware, the major difference being the replacement of the earlier underglaze blue by an overglaze blue. On most genuine examples it is possible to see a distinct halo around the overglaze blue, but its absence does not condemn the piece as not genuine. The *famille noire* has the *verte* palette in conjunction with a black ground; the *famille jaune* uses the same colours, but is used in conjunction with a yellow ground. In each case the white porcelain disappears under the colours.

During the reign of the K'ang-hsi emperor (*c.* 1685) an opaque rose-coloured overglaze appears. This and its related colours were called "foreign colours" (*yang ts'ai*). It soon formed the characteristic colour of a group of wares, referred to as the *famille rose*, which was particularly developed during the reign of the Yung-cheng emperor. It more or less replaced the *verte* palette. The translucent overglaze colours of the earlier period tended to become opaque, and painting has a more feminine quality.

During the 18th century the white wares of Ching-te-chen were made mostly for the home market, though a few were exported. They included examples of the bodiless ware and the *an hua* (literally "secret language"). The latter, copied from a traditional Yung-lo (1402–24) type, has designs lightly incised or painted with white slip. The body



Figure 140: Trumpet-shaped vase with floral decoration on background of green enamel, *famille verte*, Ch'ing dynasty, reign of the K'ang-hsi emperor, 1661–1722. In the Victoria and Albert Museum, London. Height 61 cm.

By courtesy of the Victoria and Albert Museum, London

is white, and the whole is covered with clear glaze. The decoration can only be seen plainly if light is allowed to shine through it. Pierced work was revived in certain rare pieces inspired by jade; the use of piercing that was filled with glaze was derived from Persian Gombroon ware.

European influence and the export trade. Before the mid-18th century, some European wares had found their way to China, as witness certain copies of early Meissen porcelain. The taste of the European trader, though hardly representative of the more cultured section of Western civilization, also began to have influence.

Much decoration was done in studios in and around the port of Canton, white porcelain being sent from Ching-te-chen for the purpose. Enormous quantities of *famille rose* porcelain were painted there, including most of the "ruby-backed" dishes, which are completely covered on the reverse, except for the interior of the foot ring, with a ground of overglaze opaque rose. They often have an elaborate arrangement of minutely delineated border patterns around the central subject (usually pretty women), demonstrating the new, and later widespread, idea that the beauty of an object is directly proportional to the amount of decoration on it. This theory was to be one of the causes of the degeneration of later Chinese and Japanese wares; it was, however, by no means confined to the Orient and can be seen in most 19th-century European porcelain.

The Yung-cheng painters were the first to carry foliate decoration over on to the back of the dish, usually as a prolongation of the stem. This was repeated later during the reign of the Tao-kuang emperor (1820–50). The European tendency to draw flowers in a naturalistic manner also appears in China from the Yung-chen period onward, although the practice was not carried to the same lengths.

An attempt to imitate the European method of overglaze painting, in which colours were applied in flat washes that partly sank into soft porcelain glazes, can be seen in the "ancient moon pavilion" (*ku yüeh hsüan*) wares. These will sometimes have a European subject, for example, a Watteau shepherdess, but Chinese subjects were also used.

Oriental Lowestoft

Of the wares more directly due to European intervention perhaps the best known is Chinese export porcelain, still sometimes known as Oriental Lowestoft. The name is due to an error on the part of William Chaffers (the author of a book on pottery marks), who persisted in attributing these wares to the small English factory at Lowestoft. If this porcelain is important at all, it is as a curiosity; the artistic value is nearly always negligible. The styles are usually based on those of European pottery or metalwork or on a combination of Western and Oriental motifs in an unpleasing jumble. The designs were provided by Western traders, and coats of arms are comparatively common.

Other wares connected with the export trade are those decorated with the Mandarin patterns; these came from Cantonese studios and were introduced toward the end of the 18th century. They have figure subjects in panels that are surrounded with coloured grounds and an excess of floral and other ornament in unprepossessing combinations of colours.

Much white porcelain was sent to Canton to be decorated, but much, too, was shipped to Europe for the same purpose. Many examples were painted by German studio painters, by Dutch enamellers, and by English "outside decorators."

Europe, of course, was not the only export market open to the Chinese. Much blue-and-white was exported to the traditional markets in Persia and the Near East (Arabic inscriptions can be seen on some specimens) and elsewhere—India, Thailand, Burma, and Tibet.

Other Chinese kilns

The wares discussed so far have been principally those of Ching-te-chen. Those of Te-hua in Fukien Province, however, are also important. Figures of the Buddhist goddess Kuan-yin in particular were exported in enormous quantities, and the *an hua* and pierced decorations often came from Te-hua. Vessels, such as libation cups with applied prunus sprigs, were copied by European factories in the 18th century, notably by those at Meissen, Chelsea and Bow, and Saint-Cloud. The body is usually white, sometimes with an ivory tone, and the glaze is thick, rich, and lustrous. European forms are to be seen occasionally, and most coloured examples have been decorated in the West. The kilns of I-hsing also continued making the traditional wares.

19th and 20th centuries. The 19th century has little to offer that is new or of good quality. Snuff bottles painted with miniature designs were first made toward the end of the 18th century, but most belong to the reign of the Chia-ch'ing (1796–1820) and Tao-kuang (1820–50) emperors. Bowls with circular medallions painted in overglaze colours with yellow or rose grounds are, perhaps, among the finer wares. Also of good quality are bowls covered with an opaque ground, rose or yellow, with designs engraved into it. These were first made in the 18th century and extend to the reign of the Tao-kuang emperor.

Most of the wares of Tao-kuang are poor in quality, although some examples in the style of Yung-cheng are better. The glaze has a musliny texture similar to that seen on some early Ming wares and on Japanese porcelain from Arita. Translucent overglaze colours over underglaze blue are a Yung-cheng type that had a revival at this time. In addition, the *rose-verte* palette was commonly used.

In 1853 the Taiping Rebellion led to the destruction of the kilns at Ching-te-chen, which were not rebuilt until 1864. The reign of the T'ung-chih emperor (1861–75) is principally notable for poor copies of earlier monochromes, including the peach-bloom glaze. Nearly all wares from this time onward are slick copies of older work.

KOREA

Because Korea lies to the north of China and close to the islands of Japan, it has usually formed a cultural link

between the two countries. During the Japanese invasion of 1592, for instance, many Korean potters were taken to Japan, where they were set to work making tea ceremony wares, which had hitherto been imported, and they later helped to found the porcelain industry.

It is difficult to distinguish some Korean wares from those made in the northern provinces of China during the contemporary Han to T'ang period. The wares of the Silla period (57 BC–AD 935) include some reminiscent of those of the Chou dynasty. Specimens of stoneware obviously based on metalwork are distantly related to some of the Han bronzes. Patterns on these wares are geometric and incised into the clay before firing.

An olive-green glaze was introduced later in the Silla dynasty, probably about the 9th century. Roof tiles and finials have a brown or a green glaze and may be contemporary with the Han dynasty.

The wares of the Koryŏ dynasty (918–1392; roughly corresponding to the Chinese Sung and Yüan dynasties) exhibit a much greater diversity and fall into rather more clearly defined groups. The attribution of certain black-glazed *temmoku* types (see above *China: Sung dynasty*) is controversial, but it seems that at least some of them were made in Korea. Many celadons, too, have typical Korean lobed forms, based on the melon or the gourd. These are also to be seen in porcelain, much of which has a bluish-white glaze. Some lobed boxes, usually circular, are decorated with impressed designs and are probably always Korean.

One of the difficulties in the study of Korean pottery is that practically everything has been recovered from tombs; few actual kiln sites have been discovered. Nevertheless, one such excavation at Yuch'ŏn-ni has disclosed shards of both the celadon glaze and of white porcelain from which it seems evident that white porcelain resembling both the *ying-ch'ing* and Ting types was made (see above *China: Sung dynasty*). The earliest vessels were probably fairly close copies of Chinese styles, while the distinctive Korean style followed rather later. A crazing of the glaze and a certain amount of flaking are characteristic. A mere handful of specimens, some fragmentary, of inlaid white porcelain have survived. They are best represented by a vase in the Natural Museum of Modern Arts in the Tŏksu Palace of Seoul that has panels of black-and-white inlay beneath a celadon glaze. Decoration on much Korean porcelain of the period is either incised (foliage being a frequent motif), combed, or molded in shallow relief.

White porcelain

Korean celadons have a stoneware body covered with a glaze varying from bluish green to a putty colour; some are obviously analogous to the celadons of Yüeh-chou. Characteristic of Korean pots are the stilt or spur marks to be seen on the otherwise glazed base; these are the points on which the pots rested in the kiln. Many of the forms are lobed. Perhaps the most important divergence from the usual Chinese celadon is the presence of inlaid decoration beneath the glaze of many specimens, later examples of which are often referred to as *mishima* (Figure 141).

Mishima decoration

The designs were first incised into the clay, and the incisions were then filled with black and white slip. The inlaid patterns are diverse, but most of the subjects are floral; birds are to be seen occasionally. Isolated flowers with symmetrically radiating petals are also found, principally on boxes.

While most Korean wares of the Yi dynasty (1392–1910) are distinctly rougher than those of China in the Ming and Ch'ing periods, the decoration is often magnificent in quality. Most can be clearly distinguished from Chinese wares by their forms, which show distinct differences in almost every case. Lobed forms suggested by the melon are very characteristic, and the pear-shaped bottle differs in its proportions from that of the Chinese. The large rugged jars with high shoulders are not so precisely potted as similar jars from China, often showing a marked degree of asymmetricality. Twisted rope handles are also peculiar to Korea. Many of the ewers are obvious adaptations from metalwork.

Painting in brownish black beneath a celadon glaze, which had begun in the Koryŏ dynasty, continued in the Yi dynasty. Inlaid decoration was also executed during the early part of this period, the pattern often being engraved by stamps rather than incised freehand. Sgraffito decoration, in which patterns were incised through a grayish-white slip, is also seen occasionally.



Figure 141: Korean bottle with a celadon glaze and *mishima* (inlaid decoration), Koryŏ dynasty, 13th century. In the Victoria and Albert Museum, London. Height 34.6 cm.

By courtesy of the Victoria and Albert Museum, London; photograph, Wilfrid Walter

Some excellent painted designs in an underglaze blue of variable colour but usually distinctly grayish in tone were executed on a rough porcelain body that is almost stoneware. The designs are particularly notable for great economy of brushwork and superb drawing. Their affinities are much more with Japanese pottery than with contemporary Chinese wares. A typical Japanese technique, "brush" (*hakeme*), or brushed slip, is used in conjunction with painted decoration in the early part of the dynasty, but later it is used alone. Korean influence on Japanese pottery was probably at its strongest during the ascendancy of the Japanese warrior Hideyoshi (1536–98), who invaded Korea. It is unlikely that much important work was done in Korea itself after this invasion.

JAPAN

Since Japan is a well-wooded country, wood has always been used for domestic utensils of all kinds, either in a natural state or lacquered. Until recent times, therefore, pottery and porcelain were not employed extensively for general domestic use but were reserved for such special purposes as the tea ceremony (see below). In pottery the Japanese especially admire accidental effects that resemble natural forms. Objects that appear misshapen and glazes that exhibit what would normally be regarded as serious imperfections in the West are admired by the Japanese connoisseur. The Japanese potter liked his work to reveal the impress of the hand that had made it. Marks, such as the ridges left by the fingers in a newly thrown vessel, were frequently accentuated instead of being obliterated, and marks made by tools were often left untouched.

Accidental effects

Hand modelling was practiced long after the wheel was known, and asymmetries and irregularities of form were purposely sought. Similar accidental effects were encouraged in glazing: coloured glazes were allowed to run in streaks and were irregularly applied. They were often thick, with many bubbles, and with a semifluid or treacly appearance. Crackled glazes and those deeply fissured (the latter called dragon skin or lizard skin) were deliberately

induced. Painted decoration, frequently blue, brown, or iron red, is often summary and almost calligraphic in its simplicity. The aim was to give an overall effect that resembled such natural objects as stones, in being largely uncontrived.

From the 15th century onward, the art of the potter was also affected by the elaborate tea ceremony (the *cha-no-yu*). In its original form it was probably introduced from China by Zen priests, but at the court of the shogun (military governor) Yoshimasa (1435–90), in Kyōto, it developed into a fixed ceremonial pattern. Possibly the ceremony was first exploited as a means of settling feudal disputes. It is held in a small room or pavilion, usually surrounded by a carefully designed garden. When the guests are summoned they enter a sparsely furnished room through a very low doorway. The fact that guests must crawl into the room is thought to have served the purpose of preventing them from concealing a sword under their robes.

In a recess called the toko-no-ma, a picture mounted on brocade or silk is hung, and the guests bow to this in appreciation. The tea master puts a little powdered tea in a bowl and pours on it water that has been heated over a charcoal brazier. The tea is whipped to a froth with a bamboo whisk and then passed from hand to hand. The various utensils (the teabowl, tea caddy, water container, boxes, plates, and iron tea kettle) have been carefully selected by the tea master and are often of great age. The tea drinking is followed by a discussion and appreciation of the qualities of the utensils. The bowls are valued for their heat-retaining properties and the way in which they fit the hand as well as for their appearance. Sometimes a newly acquired work of art is produced by the host for the delectation of his guests. Since the tea masters were the aesthetic arbiters, their influence on Japanese pottery was profound.

Jōmon pottery

The early history of Japan is considerably more obscure than that of China. The first Japanese pottery belongs to the Jōmon period (*c.* 2500–250 BC). It has a black body, and the decoration is usually an impressed representation of coiled rope or matting (*jōmon* means "coiled"). *Jōmon-shiki* ("pottery") is widely distributed throughout the islands, but complete specimens are very rare. It was followed by Yayoi pottery, specimens of which have been excavated throughout Japan. The body is somewhat finer in quality than Jōmon pottery and is usually red or gray. Decoration is simple, and forms will sometimes show the influence of Korean pottery of the period. It ceased to be produced about the 6th century AD.

Meanwhile, from about the 3rd to the 6th century AD, large tombs were constructed in the form of oval or circular tumuli from whose bases have been recovered the *haniwa* ("clay circle") figures of warriors, women, horses, and so forth. They are hollow, and, though vigorously modelled, they are more primitive than analogous tomb figures from China.

In the Asuka or Sueki period (AD 552–710) that followed, wares are much more sophisticated. Unlike the preceding types, they were made with a wheel, and firing took place in a rudimentary kiln at a much higher temperature than previously. Widespread manufacture continued through the Nara period (710–794) and the early part of the following Heian, or Fujiwara, period (794–1185). Some examples have a smear glaze, no doubt at first caused accidentally by wood ashes coming into contact with the surface. Three colours of glaze—green, yellowish brown, and white—were used either alone or in combination and resemble those of T'ang earthenware. Pottery of this kind has been found around Ōsaka and Kyōto. The principal pottery productions of the period were vases, dishes, bowls, and bottles of various descriptions.

The influence of Korea and of T'ang China is noticeable. Toward the end of the Heian period contacts with China were severed, and there was a corresponding decline in the art of pottery; even the traditional Sueki ware disappeared.

Kamakura and Muromachi periods (1192–1573). A revival in the Kamakura period (1192–1333) followed the visit of the potter Katō Shirōzaemon (Tōshirō) to China in 1227, where he learned the secrets of pottery making. He established himself at Seto, Owari (now Aichi Prefecture), which speedily became a large centre of manufacture. There were soon about 200 kilns in the vicinity making a variety of wares, some of which were glazed in black in imitation of the *temmoku* wares of China (see above *China: Sung dynasty*). The early wares were mainly for ritual purposes, but by the beginning of the Muromachi, or Ashikaga, period (1338–1573) teabowls, plates, jars, and saucers of domestic utility were also being made. Wares of the Kamakura period are decorated with incised designs or with impressed or applied ornament. The Muromachi wares are much plainer as the result of the growing influence of the tea ceremony, especially the *wabi* school of the cult, which concentrated on rustic simplicity. The wares of both of these periods have a feldspathic glaze, but the Muromachi glaze is more even in quality than the Kamakura, which has a tendency to run in rivulets. A transitional type has a soft-yellowish glaze or a dark-brown glaze sometimes called Seto *temmoku*.

A large number of kilns were in existence, the more important known as the "six pottery centres of ancient Japan." These were Seto; Tokoname (also in Aichi Prefecture), which may have exceeded Seto in the size of its production; Bizen (Okayama Prefecture), which has produced an excellent unglazed stoneware from the Heian period to the 20th century; Tamba (Kyōto Prefecture); Shigaraki (Shiga Prefecture); and Echizen (Fukui Prefecture). The wares of Seto, especially those made for Buddhist ceremonies, were regarded as the finest pottery of this period.

Seto–Mino wares

Azuchi-Momoyama period (1573–1600). Production had been interrupted during the civil wars of the 15th and 16th centuries. Toward the end of the 16th century the Seto kilns were removed for a time to the Gifu Prefecture of Mino Province, where they received the protection of the feudal baron (daimyo) of Toki. The Mino pottery was founded by Katō Yosabei, whose sons started other potteries in the vicinity, notably that under the aegis of the tea master Furuta Oribe Masashige. New kilns were also built elsewhere, and pottery, while retaining its importance in the tea ceremony, became much more widely used for ordinary purposes. The inspiration for most of its shapes and designs came from the Mino region. The later wares of these kilns are much less austere than those attributed to the Muromachi period, since the cult of the tea ceremony, now widespread, had lost something of its earlier simplicity. Characteristic tea ceremony wares of the early years of the 17th century are Shino, which has a thick, crackled glaze and is sometimes summarily painted in blue or brown; yellow Seto (*ki-Seto*), whose crackled yellow glaze covers a stoneware body; and, at Narumi, in the adjoining Owari region, a ware of the kind associated with Oribe (which had become a generic term for pottery influenced by the tea master of that name), which is glazed in white, straw colour, yellowish green, and pinkish red, with sometimes the addition of slight painting in brown.

Toward the end of the 16th century the tea ceremony was reformed by Sen Rikyū (1521–91), the tea master to the military dictator of Japan, Toyotomi Hideyoshi. Sen Rikyū was principally responsible for the replacement of the hitherto much admired *temmoku* bowls from China by others patterned after unsophisticated Korean wares; his influence has persisted to the 20th century. In the 1590s Hideyoshi twice invaded Korea, and as a result of these wars many Korean potters were taken to Japan, where their influence was considerable.

Raku ware

A tilemaker named Ameya, who is said to have been a Korean, introduced a type of ware that was covered with a lead glaze and fired at a comparatively low temperature. His son Tanaka Chōjirō and his family extended this technique to the tea bowl, and in about 1588 their wares were brought to the notice of Hideyoshi, who awarded them a gold seal engraved with the word *raku* ("felicity"). The *raku* (Figure 142), made in Kyōto, are among the most famous of all Japanese wares. The shape of the vessels is extremely simple: a wide straight-sided bowl set on a narrow base. At first the glaze was dark brown, but a light orange red was developed later, to be followed in the 17th century by a straw colour. Still later, green and cream and other colours were introduced. Tea-bowls attributed to the first Chōjirō are greatly valued in Japan.



Figure 142: *Raku* ware water jar and cover, Tokyo, Edo period, 18th century. In the Museum of Fine Arts, Boston. Height 20.3 cm.
By courtesy of the Museum of Fine Arts, Boston

Karatsu ware

The kilns of Karatsu, a district in the north of Hizen Province, may have been established by Korean potters, since the influence of Korea is perceptible in some of them. The term Karatsu ware encompasses a great variety of shapes and styles: "undecorated" (*muji*), "painted" (*e*), "speckled" (*madara*), in the Korean style (Chosen), which has a thick opaque glaze, and in the style of Seto, which has a white glaze. The earliest Karatsu ware belongs probably to the end of the 16th century, although it is sometimes awarded a still earlier date. Most surviving examples belong to the 17th century. The most valued pieces are those made for the tea ceremony. (Ge.S.)

Edo period (1603–1867). According to tradition, the first Japanese porcelain was made in the early 16th century after Shonzui Goradoyu-go brought back the secret

By courtesy of the Metropolitan Museum of Art, New York, Rogers Fund, 1918



Figure 143: Kakiemon ware decorated with iron red and coloured enamels, Hizen Province, Japan, Edo period, 17th century. In the Metropolitan Museum of Art, New York. Diameter 24.4 cm.

of its manufacture from the Chinese kilns at Ching-te-chen. Another account claims that Ri Sampei, a Korean potter who was brought to Japan by Hideyoshi, discovered porcelain clay in the Izumi Mountain near Arita (Saga Prefecture); this version is feasible since no porcelain made before the end of the 16th century has been identified.

The first Arita manufacture was decorated in blue underglaze, simple and excellent in quality. Specimens soon found their way to Europe in Dutch ships, and the Dutch were awarded a trading monopoly in 1641. Some of these early Japanese export wares are based on contemporary European metalwork and faience.

Kakiemon porcelain

The family of Sakaida is especially connected with the Arita kilns. The first recorded member, born about 1596, worked in underglaze blue until the family learned the secret of using overglaze colours. According to tradition, it was told to them by a Chinese met by chance in the port of Nagasaki. This overglaze technique was perfected soon after the middle of the 17th century. It was continued by the family, and, since many of them were called Kakiemon, the style has become known by that name. The palette is easily recognized—iron red, bluish green, light blue, yellow, and sometimes a little gilding; many examples have a chocolate-brown rim. Octagonal and square shapes are especially frequent (Figure 143). Themes of decoration are markedly asymmetrical, with much of the white porcelain surface left untouched. This technique and style spread rapidly to other provinces, and its influence on porcelain that was manufactured in Europe during the first half of the 18th century was at least as great as that of Chinese porcelain. A later Kakiemon development in which "brocade" (*nishikide*) patterns in compartments were used (at the suggestion of Dutch traders) proves to be less pleasing. These later coloured wares from Arita became known as Imari, after the port from which they were shipped.

Like 18th-century Chinese white porcelain, Japanese white wares were shipped to Europe, where they were decorated by Dutch and other European enamellers.

Of considerable importance but more rarely seen in Europe is the porcelain called Kutani. The kiln at Kutani in Kaga Province (now in Ishikawa Prefecture) operated in the latter half of the 17th century. Greatly valued, Old Kutani (*ko*-Kutani) porcelain is among the finest of the Japanese wares. The body is heavy, approaching stoneware, and the designs are executed boldly and in rich colours. Old Kutani was revived and other styles arose when kilns in the area resumed operation in the early 19th century and again in the 1860s, the latter resulting in the establishment of modern "Kutani ware" as a major export item.

The Mikawachi kilns under the protection of the prince of Hirado made porcelain principally for his use. The delicate, very white body is usually decorated in miniature style with underglaze blue. Kyōto imitated Sung celadons and the Ming green and red wares. Seto made no porcelain until about 1807; the first production was decorated in underglaze blue (*sometsuke*). Overglaze colours date from about 1835.

The manufacture of earthenware was continued during the 17th and 18th centuries, and much of it is notable for its decoration. Toward the end of the 17th century, Ninsei (Nonomura Seisuke) began work at Kyōto and was responsible for much finely enamelled decoration on a cream earthenware body covered with a finely crackled glaze. Also produced at Kyōto, the works of Kenzan, who used rich and subtly coloured slips often as a background for plant motives, and of the Dōhachi family, famous for their overglaze decoration, are much sought after in Japan.

Satsuma pottery

19th and 20th centuries. Japanese productions during the 19th century, in common with those in most other parts of the world, greatly deteriorated in taste. Typical of the period is the so-called Satsuma pottery, most of which was made not at Satsuma but at Kyōto and then sent to Tokyo to be decorated especially for export. The designs are overcrowded and debased, and its popularity undoubtedly retarded an appreciation of work in the true Japanese taste among Western students and collectors.

Like Western pottery manufacture in the mid-20th century, Japan's is largely industrialized, and most products are derivative, but the Japanese tradition of pottery making in small and private kilns continues. (Ge.S./Ed.)

THAILAND AND ANNAM

Pottery was made in the old Siamese capitals of Sukhothai and Sawankhalok. It is also thought that potteries persisted at Ayutthaya until the 18th century. Little is known of the early history of the region, and definite information on its pottery is almost nonexistent. Dating of the pottery from these regions for the most part has been by ana-

logy with related Chinese wares, which greatly influenced Siamese work.

Kilns have been excavated on the site of old Sawankhalok, about 200 miles (320 kilometres) north of Bangkok. The principal type of ware is a grayish-white stoneware covered with a translucent celadon glaze, usually grayish green in colour. The glaze is commonly crackled; this appears to be fortuitous, since little trouble was taken to achieve a precise finish. A particularly common decoration consists of roughly scored vertical flutes, with incised circles at the shoulder to accentuate the form. Decoration of a more definite kind is always incised under the glaze and is usually floral. Flowers are stylized, sometimes with combed lines on the petals. Covered bowls, dishes, ewers, and bottles with two small loop handles at the neck are the most common forms.

Another type of ware, with similar forms, has a rather treacly-brown glaze. Some well-modelled animals are found with this glaze. There are also tiles and bricks with crudely modelled figures in relief on them. They are analogous in form and technique to Chinese pottery of the Sung dynasty and are generally regarded as being contemporary with the Sung or Yüan period. Some small covered jars of a gray porcelaneous ware, summarily decorated with stylized floral and foliate patterns, appear to have been made at Sawankhalok (the date is probably equivalent to that of the early Ming period). These Martabani wares were widely exported throughout the East during this period.

Little is known of wares made in Annam, but some brownish celadons are regarded as likely to have been made there, as well as some small covered jars painted in a poor underglaze blue.

American Indian pottery

The American Indians are of Asiatic descent; their route to the New World was from Siberia into Alaska across the Bering Straits. The usually quoted period of their migration is between 40,000 and 10,000 years ago. Since they were nomadic peoples, it is unlikely that they brought the knowledge of pottery making with them. When pottery making did begin, it was fundamentally unlike any known work from the Old World, and the few remote resemblances to Oriental motifs are almost certainly fortuitous. The wheel remained unknown until the arrival of Europeans, although there is reason to think that a turntable, or slow wheel, may have been used occasionally. Most of the pottery was made by coiling, some by molding—both are techniques that could have arisen spontaneously. It is likely that most of the work was done by women rather than by the men. This is nearly always the case with primitive potters when the wheel is not used, and Pueblo Indian women still do this kind of work.

Slips were used to cover the body, and coloured slips provided the material for much of the painted freehand decoration. Glazes are rare, although examples can be found among the Pueblo Indians of New Mexico from about AD 1300 onward, on a few vessels from the Chimú area in the Andes, and occasionally in Central America. The effect of a reducing atmosphere was understood, so that gray and black pots are found as well as the red and brown ones fired in an oxidizing flame. Undecorated surfaces were often highly polished.

NORTH AMERICA

The most important North American pottery was made in the southwest—an area including Arizona, New Mexico, and also parts of Utah and Colorado. The people who inhabited the plateau land from about 100 BC are often referred to as the Anasazi, a Navaho Indian word meaning ancient people. They are the ancestors of the Pueblo, who began to emerge about AD 700. The Anasazi were nomadic hunters; although they did not at first make pottery, they did make excellent baskets. Fixed dwellings appear about AD 50, and this probably marks the beginning of pottery manufacture. The earliest pots appear to have been baskets that were smeared with clay and then dried in the sun.

Next came basket-shaped wares coiled in a gray body, used principally for cooking. They were followed by more decorative bowls and pots, with striking black and white geometric designs that seem to have been executed about AD 700. Slightly later there is another type of ware that has black decoration on a red slip. After the 12th century the earlier types began to disappear and were replaced by polychrome wares decorated with stylized birds, feathers, animals, and human figures amid the geometric patterns. The principal colours are yellow and red. A small quantity of glazed ware was made in the Zuñi area of New Mexico.

The Hohokam tribes (a Pima word meaning "those who have gone"), who lived in the desert of southern Arizona and were approximately contemporary with the Anasazi, made pottery figures for religious purposes, usually of crudely modelled naked women. Some of this pottery is a gray ware, but most of it is buff, with decoration in iron red that has a quality lacking the stiffness of the Pueblo designs.

Pottery figures for religious purposes

The Mogollon culture of New Mexico produced, during the Mimbres period of the 11th and 12th centuries, a ware remarkable for its lively black and white decoration depicting human, animal, and insect forms in a much less stylized manner than the paintings on most other wares from the southwest.

There is little pottery of importance from other parts of the United States. Primitive pots have been found on the Atlantic coast, in Georgia and Florida, on the Gulf Coast, and elsewhere, some of which are based on basketwork. Geometric decoration, usually incised, is the rule. Eskimo pottery, which is generally rather crude, bears some resemblance to early Asiatic types.

CENTRAL AMERICA

The pottery of Mexico and the rest of Central America is of considerable interest, but the wares are so diverse that it is impossible to summarize them adequately. They probably date from the 2nd millennium BC onward and were made by the Mayas, the Zapotecs, the Toltecs, and the Aztecs. Generally speaking, geometric patterns are common, and slips in black, brown, white, or red were frequently used. A curiosity of Central America (possibly adopted from South America) is a technique that resembles to some degree the batik method of dyeing textiles.

The surface of the pot was coated either with wax or gum. This was then scraped away in part to form a predetermined pattern, and the whole surface of the pot was covered with pigment. In firing, the gum burned away, leaving only the scraped parts in colour. Ornament carved in low relief after firing is to be seen occasionally and has few parallels outside the Americas. An unusual tech-

Maya, Zapotec, Toltec, and Aztec pottery

By courtesy of the trustees of the British Museum



Figure 144: Stirrup spout vessel in the form of a portrait head, Mochica culture, Peru (100 BC–AD 900). In the British Museum. Height 21.6 cm.

nique from the Mexican highlands consisted of covering the whole surface of the pot with a kind of thick slip, most of which was then scraped away, leaving only thin partitions. These compartments were filled with slips of a contrasting colour. The commonest shapes are bowls and wide-mouthed vases; many of these were made with legs, usually three, so that they could be set down on uneven ground. Figurines, some of which are painted, have also been found.

Between about 600 BC and AD 1000, the Mayas were making an excellent polychrome pottery in which designs in red and black were painted on a cream or orange slip. Between roughly the 4th and the 10th centuries AD, the Zapotecs, whose chief ceremonial site was Monte Albán in Oaxaca State, made striking urns in the forms of their gods. An orange-coloured pottery decorated in a great diversity of styles is associated with the Toltecs, as is a dark-coloured pottery with a glossy appearance and incised ornament (plumbate ware). Both of these types were widely distributed throughout Central America from the 11th to 14th centuries. Little is known of the Aztecs until about 1325, the date of the foundation of Tenochtitlán (Mexico City). Much of their later pottery utilizes an orange-burning clay that was painted with black curvilinear geometric motifs, in contrast to their earlier rectilinear style. During the period of Montezuma I in the 15th century, designs became more naturalistic, and birds, fish, and plant forms were freely utilized. European motifs first appear after the conquest, and such techniques as tin glazing were used from the 17th century onward.

SOUTH AMERICA

Most South American pottery was made at centres in the Andes and on the west coast, particularly in Bolivia and Peru. Pottery of lesser importance comes from Ecuador, Colombia, northwest Argentina, and northern Chile. In some places a very high degree of skill was attained, especially in the central Andes, where the earliest wares seem to date from the end of the 2nd millennium BC. Much of the pottery was made in molds. The stirrup-shaped spout on many jars is a characteristic feature. The batik type of decoration already mentioned was also used. Vessels were modelled in the shape of animal or human figures, which were also used as motifs for painted decoration. The puma god worshipped by the early peoples appears in many forms. Depictions of erotic themes also appear in South American pottery, particularly in the Mochica and Chimú cultures.

Mochica culture

The work of the Mochica culture, which flourished around the northern coast of Peru, is at its best about the 7th century AD. Jars in the form of human heads, some of which may be portraits, are remarkable both for the naturalism of the treatment and the skill of the potter. These have the stirrup spout (Figure 144). Painted decoration is often stylized although with a considerable degree of realism, and the subjects are nearly always ceremonial or religious nature.

The pottery of the Nazcas, who lived on the southern coast of Peru at much the same period, is noted for its painting. A varied palette included several shades of red and blue, yellow, orange, green, brown, black, gray, and white. The stirrup spout here becomes two spouts joined by a flat bridge. The earliest painting is on a red ground, white grounds becoming more common later. Geometric patterns are to be seen in conjunction with stylized birds, human heads, and the like. The naturalistic portrait jars of the Mochica do not appear, but there are some vessels in the form of figures modelled in a much more conventional style and similarly painted. Puma's heads occur in relief, with the body of the animal completed in brushwork. The centipede god is a motif that does not appear elsewhere.

The people of Tiahuanaco, who lived in the region around Lake Titicaca, were influenced by the Nazca wares, though painted decoration, often carried out on a red slip ground, is more limited in colour than the Nazca. The puma head was used as a motif. In general, decoration is extremely stylized with a very strong geometric flavour.

Chimú culture

The Chimú culture succeeded the Mochica in the northern area and lasted until the arrival of the Incas. The most familiar ware is in a body that varies from gray to black, although a red polished ware, sometimes painted in white slip, was also made. The influence of the Mochica tradition can be seen in the retention of the stirrup spout on some jars; others have the double spout connected by a flat bridge. The modelled wares of the Mochica culture were also revived but are of a generally inferior quality.

The Incas originally settled in Cuzco, the old capital of Peru, at the end of the 11th century. During the 15th century they established themselves over a wide area, including the territory of the Chimús. They were principally soldiers and administrators with small inclination toward luxury; and their pottery, of excellent quality particularly in the 15th century, is designed without an excess of decoration. Most Inca pottery is red polished ware. It is usually painted with geometric designs in red, white, and black, although relief decoration is also seen on black ware, especially from the Chimú region.

The commonest surviving form has been called an aryballos, although its resemblance to the Greek form is remote and fortuitous. It has a conical base, and the neck finishes in a flaring mouth. Two loop handles are set low on the body. The assumption that this vessel was made for carrying water on the back seems a little doubtful in view of its shape and the disposition of the handles. Little fine pottery was made after the arrival of the Spaniards in the 16th century. (Ge.S.)

BASKETRY

Basketry in myth

The Babylonian god Marduk "plaited a wicker hurdle on the surface of the waters. He created dust and spread it on the hurdle." Thus ancient Mesopotamian myth describes the creation of the earth using a reed mat. Many other creation myths place basketry among the first of the arts given to man. The Dogon of West Africa tell how their first ancestor received a square-bottomed basket with a round mouth like those still used there in the 20th century. This basket, upended, served him as a model on which to erect a world system with a circular base representing the sun and a square terrace representing the sky.

Like the decorative motifs of any other art form, the geometric, stylized shapes may represent natural or supernatural objects, such as the snakes and pigeon eyes of Borneo, and the kachina (deified ancestral spirit), clouds, and rainbows of the Hopi Indians of Arizona. However, the fact that these motifs are given a name does not always mean they have symbolic significance or express religious ideas.

Sometimes symbolism is associated with the basket itself. Among the Guayaki indians of eastern Paraguay, for example, it is identified with the female: the men are hunters, the women are bearers as they wander through the forest; when a woman dies, her last burden basket is ritually burned and thus dies with her.

Though it would appear that basketry might best be defined as the art or craft of making baskets, the fact is that the name is one of those the limits of which seem increasingly imprecise the more one tries to grasp it. The category basket may include receptacles made of interwoven, rather rigid material, but it may also include pliant sacks made of a mesh indistinguishable from netting—or garments or pieces of furniture made of the same materials and using the same processes as classical basketmaking. In fact, neither function nor appearance nor material nor mode of construction are of themselves sufficient to delimit the field of what common sense nevertheless recognizes as basketry.

In this section the word will be taken to mean a handmade assemblage of vegetable fibres that is relatively large and rigid, so as to make a continuous surface, usually (but not exclusively) a receptacle. The consistency of the mate-

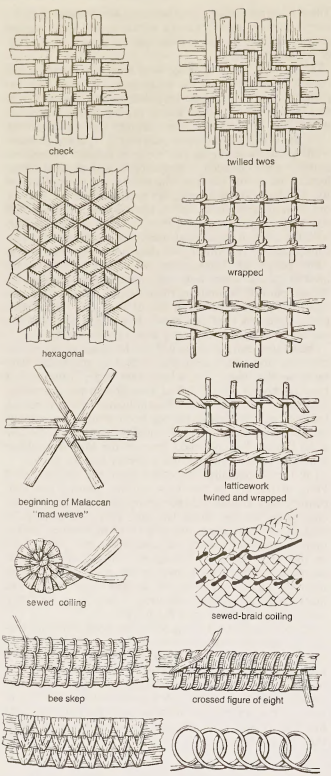


Figure 145: Varieties of plaited and coiled work used in basketry.

rials used distinguishes basketry, which is handmade, from weaving, in which the flexibility of the threads requires the use of an apparatus to put tension on the warp, the lengthwise threads. What basketry has in common with weaving is that both are means of assembling separate fibres by twisting them together in various ways.

This section deals with the construction materials and techniques of basketry as well as their symbolism, uses, and history.

Materials and techniques

There is no region in the world, except in the northernmost and southernmost parts, where people do not have at their disposal materials—such as twigs, roots, canes, and grasses—that lend themselves to the construction of baskets. The variety and quality of materials available in a particular region bears on the relative importance of basketry in a culture and on the types of basketry produced by the culture. Rainy, tropical zones, for example, have palms and large leaves that require plaiting techniques different from those required for the grass stalks that predominate in the dry, subtropical savanna regions or for the roots and stalks found in cold temperate zones. The interrelationship between materials and methods of construction might in part explain why the principal types of basketry are distributed in large areas that perhaps correspond to climatic zones as much as to cultural groups: the predominance of sewed coiling, for example, in the African savannas and in the arid zones of southern Eurasia and of North America; of spiral coiling and twining in temperate regions; and of various forms of plaiting in hot regions. There is also a connection between the materials used and the function of the basket, which determines whether rigid or soft materials—either as found in nature or specially prepared—are used. In East Asia, for example, twined basketry fashioned out of thin, narrow strips (called laths) of bamboo is effective for such objects as cages and fish traps that require solid partitions with openings at regular intervals. Soft and rigid fibres are often used together: the rigid fibres provide the shape of the object and soft ones act as a binder to hold the shape.

Aesthetic goals and choice of material

Finally, materials are chosen with a view toward achieving certain aesthetic goals; conversely, these aesthetic goals are limited by the materials available to the basket maker. The effects most commonly sought in a finished product are delicacy and regularity of the threads; a smooth, glossy surface or a dull, rough surface; and colour, whether natural or dyed. Striking effects can be achieved from the contrast between threads that are light and dark, broad and narrow, dull and shiny—contrasts that complement either the regularity or the decorative motifs obtained by the intricate work of plaiting.

Despite an appearance of almost infinite variety, the techniques of basketry can be grouped into several general types according to how the elements making up the foundation (the standards, which are analogous to the warp of cloth) are arranged and how the moving element (the thread) holds the standards by intertwining among them (Figure 145).

Coiled construction. The distinctive feature of this type of basketry is its foundation, which is made up of a single element, or standard, that is wound in a continuous spiral around itself. The coils are kept in place by the thread, the work being done stitch by stitch and coil by coil. Variations within this type are defined by the method of sewing, as well as by the nature of the coil, which largely determines the type of stitch.

Spiral coiling. The most common form is spiral coiling (Figure 146), in which the nature of the standard introduces two main subvariations: when it is solid, made up of a single whole stem, the thread must squeeze the two coils together binding each to the preceding one (giving a diagonal, or twilled, effect); with a double or triple standard the thread catches in each stitch one of the standards of

By courtesy of H. Baffot

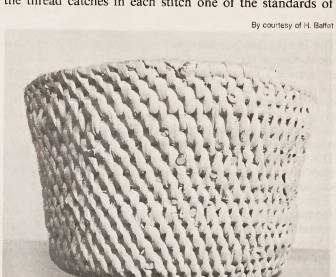


Figure 146: Spiral-coiled basket with twill effect, from Bialystok region, Poland. In the Musée de l'Homme, Paris.

the preceding coil. Many other variations of spiral coiling are possible. Distribution of this type of basketry construction extends in a band across northern Eurasia and into northwest North America; it is also found in the southern Pacific region (China and Melanesia) and, infrequently, in Africa (Rhodesia).

Sewed coiling. Sewed coiling (Figure 147) has a foundation of multiple elements—a bundle of fine fibres. Sewing is done with a needle or an awl, which binds each coil to the preceding one by piercing it through with the thread. The appearance varies according to whether the thread conceals the foundation or not (bee-skep variety) or goes through the centre of the corresponding stitch on the preceding coil (split stitch, or furcate). This sewed type of coiled ware has a very wide distribution: it is almost the exclusive form in many regions of North and West

Distribution of sewed coiling

By courtesy of Mrs. A.E. Robinson



Figure 147: Sewed-coiling basket in the tarantula design, Papago culture, Arizona. In the collection of A.E. Robinson.

Africa; it existed in ancient Egypt and occurs today in Arabia and throughout the Mediterranean basin as far as western Europe; it also occurs in North America, in India, and sporadically in the Asiatic Pacific. A variety of sewed coiling, made from a long braid sewed in a spiral, has been found throughout North Africa since ancient Egyptian times.

Half-hitch and knotted coiling. In half-hitch coiling, the thread forms half hitches (simple knots) holding the coils in place, the standard serving only as a support. There is a relationship between half-hitch coiling and the half-hitch net (without a foundation), the distribution of which is much more extensive. The half-hitch type of basketry appears to be limited to Australia, Tasmania, Tierra del Fuego in South America, and Pygmy territory in Africa. In knotted coiling, the thread forms knots around two successive rows of standards; many varieties can be noted in the Congo, in Indonesia, and among the Basket Makers, an ancient culture of the plateau area of southwestern United States, centred in parts of Arizona, New Mexico, Colorado, and Utah.

The half-hitch and knotted-coiling types of basketry each have a single element variety in which there is no foundation, the thread forming a spiral by itself analogous to the movement of the foundation in the usual type. An openwork variety of the single element half hitch (called cycloid coiling) comes from the Malay area; and knotted single-element basketry, from Tierra del Fuego and New Guinea.

Noncoiled construction. Compared to the coiled techniques, all other types of basketry have a certain unity of construction: the standards form a foundation that is set up when the work is begun and that predetermines the shape and dimensions of the finished article. Nevertheless, if one considers the part played by the standards and the threads, respectively, most noncoiled basketry can be divided into three main groups.

Wattle construction. A single layer of rigid, passive, parallel standards is held together by flexible threads in one of three ways, each representing a different subtype. (1) The bound, or wrapped, type, which is not very elaborate, has a widespread distribution, being used for burden baskets in the Andaman Islands in the Bay of Bengal, for poultry cages in different parts of Africa and the Near East, and for small crude baskets in Tierra del Fuego. (2) In the twined type, the threads are twisted in twos or threes, two or three strands twining around the standards and enclosing them. The twining may be close or openwork or may combine tight standards and spaced threads. Close twining mainly occurs in three zones: Central Africa, Australia, and western North America, where there are a number of variations such as twilled and braided twining and zigzag or honeycomb twining. The openwork subtype is found almost universally because it provides a perfect solution to the problem of maintaining rigid standards with even spacing for fish traps and hurdles (portable panels used for enclosing land or livestock). Using spaced threads, this subtype is also used for flexible basketry among the Ainu of northern Japan and the Kuril Islands and sporadically throughout the northern Pacific. (3) The woven type (Figure 148), sometimes termed wickerwork, is made of stiff standards interwoven with flexible threads. It is the type most commonly found in European and African basketry and is found sporadically in North and South America and in Near and Far Eastern Asia.

Lattice construction. In lattice construction a frame made of two or three layers of passive standards is bound together by wrapping the intersections with a thread. The ways of intertwining hardly vary at all and the commonest is also the simplest: the threads are wrapped in a spiral around two layers of standards. This method is widely used throughout the world in making strong, fairly rigid objects for daily use: partitions for dwellings, baskets to be carried on the back, cages, and fish traps (with a Mediterranean variety composed of three layers of standards and a knotted thread). The same method, moreover, can be adapted for decorative purposes, with threads—often of different colours—to form a variety of motifs similar to embroidery. This kind of lattice construction appears mainly among the Makah Indians of the U.S. Pacific Northwest and in Central and East Africa.

Uses of lattice construction

Matting or plaited construction. Standards and threads are indistinguishable in matting or plaited construction; they are either parallel and perpendicular to the edge (straight basketry) or oblique (diagonal basketry) (Figure 149). Such basketry is closest to textile weaving. The mate-

By courtesy of the University Museum of Archaeology, Cambridge, England

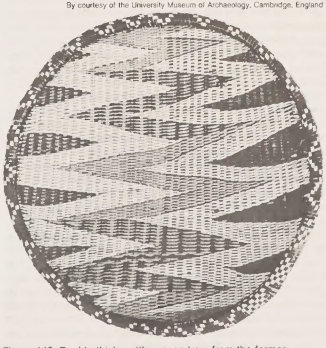


Figure 148: Double-thick wattle-woven tray, from the former Ruanda-Urundi, Africa. In the University Museum of Archaeology, Cambridge, England.

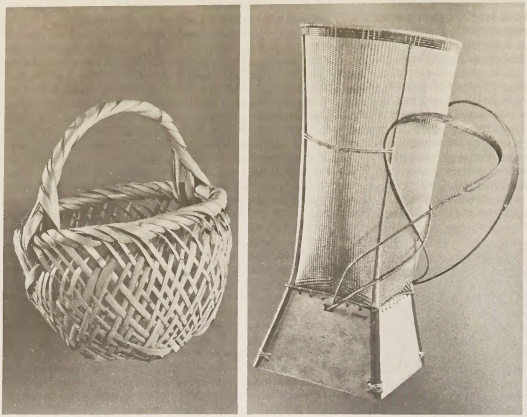


Figure 149: *Plaiting.*
(Left) Bamboo flower basket of diagonal openwork plaiting, from Japan. (Right) Burden basket of straight-woven plaiting, from Vietnam. In the Musée de l'Homme, Paris.
By courtesy of the Musee de l'Homme, Paris

rials used are almost always woven, using the whole gamut of weaving techniques (check, twill, satin, and innumerable decorative combinations). Depending on the material and on the technique used, this type of construction lends itself to a wide variety of forms, in particular to the finest tiny boxes and to the most artistic large plane surfaces. It is widely distributed but seems particularly well adapted to the natural resources and to the kind of life found in intertropical areas. The regions where it is most common are different from, and complementary with, those specializing in coiled and twined ware; that is, eastern and southeastern Asia (from Japan to Malaysia and Indonesia), tropical America, and the island of Madagascar off the east coast of Africa.

One variety of matting or plaited work consists of three or four layers of elements, which are in some cases completely woven and in others form an intermediate stage between woven and lattice basketry. The intermediate type (with two layered elements, one woven) is known as hexagonal openwork and is the technique most common in openwork basketry using flat elements. It has a very wide distribution: from Europe to Japan, southern Asia, Central Africa, and the tropical Americas. A closely woven fabric in three layers, forming a six-pointed star design, is found on a small scale in Indonesia and Malaysia.

Decorative devices. Clearly, a variety of decorative possibilities arise from the actual work of constructing basketry. These, combined with the possible contrasts of colour and texture, would seem to provide extensive decorative possibilities. Each particular type of basketry, however, imposes certain limitations, which may lead to convergent effects: hexagonal openwork, for example, forms the same pattern the world over, just as twilled weaving forms the same chevrons (vertical or horizontal). Each type, also, allows a certain range of freedom in the decoration within the basic restrictions imposed by the rigidity of the interlaced threads, which tends to impose geometric designs or at least to geometrize the motifs. In general, the two main types of basketry—plaited and coiled—lend themselves to two different kinds of decoration. Coiled basketry lends itself to radiating designs, generally star- or flower-shaped compositions or whirling designs sweeping from the centre to the outer edge. Plaited basketry, whether diagonal or straight, lends itself to over-all compositions of horizontal stripes and, in the detail, to intertwined shapes that result from the way two series of threads, usually in contrasting colours, appear alternately on the surface of the basket.

Typical coiled and plaited basketry decorations

Other art forms have been influenced ornamentally by basketry's plaited shapes and characteristic motifs. Because of their intrinsic decorative value—and not because the medium dictates it—these shapes and motifs have been reproduced in such materials as wood, metal, and clay. Some notable examples are the interlacing decorations carved on wood in the Central African Congo; basketry motifs engraved into metalwork and set off with inlayed silver by Frankish artisans in the Merovingian period (6th to 8th century); and osier patterns (molded basketwork designs) developed in 18th-century Europe to decorate porcelain.

Uses

Household basketry objects consist primarily of receptacles for preparing and serving food and vary widely in dimension, shape, and watertightness. Baskets are used the world over for serving dry food, such as fruit and bread, and they are also used as plates and bowls. Sometimes—if made waterproof by a special coating or by particularly close plaiting—they are used as containers for liquids. Such receptacles are found in various parts of Europe and Africa (Chad, Rwanda, Ethiopia) and among several groups of North American Indians. By dropping hot stones into the liquid, the Hupa Indians of northwestern California even boil water or food in baskets.

Waterproofing

Openwork, which is permeable and can be made with mesh of various sizes, is used for such utensils as sieves, strainers, and filters. Such basketry objects are used in the most primitive cultures as well as in the most modern (the tea strainers used in Japan, for example). The flexibility of work done on the diagonal is put to particularly ingenious use by the Africans in beer making and, above all, by Amazonian Indians in extracting the toxic juices from manioc pulp (a long basketwork cylinder is pulled down at the bottom by ballasting and, as it gets longer, compresses the pulp with which it had previously been filled).

Finally, basketry plays an important part as storage containers. For personal possessions, there are baskets, boxes, and cases of all kinds—nested boxes from Madagascar, for example, which are made in a graduated series so that they fit snugly one within another, or caskets with multiple compartments from Indonesia. For provisions, there

are baskets in various sizes that can be hung up out of the reach of predators, and there are baskets so large that they are used as granaries. In The Sudan in Africa, as in southern Europe, these are usually raised off the ground on a platform and sheltered by a large roof or stored in the house, particularly in Mediterranean regions; for preserving cereals they are sometimes caulked with clay.

Some of these granaries are not far from being houses. Basketry used in house construction, however, usually consists of separately made elements that are later assembled; partitions of varying degrees of rigidity used as walls or to fence in an enclosure; roofs made of great basketry cones (in Chad, for example); and, above all, mats, which have numerous uses in the actual construction as well as in the equipping of a house. Probably the oldest evidence of basketry is the mud impressions of woven mats that covered the floors of houses in the Neolithic (*c.* 7000 BC) village of Jarmo in northern Iraq. Mats were used in ancient Egypt to cover floors and walls and were also rolled up and unrolled in front of doorways, as is shown by stone replicas decorating the doorways of tombs dating from the Old Kingdom, *c.* 2686–2160 BC. It is known from paintings that they were made of palm leaves and were decorated with polychrome (multicoloured) stripes, much like the mats found in Africa and the Near East.

Woven mats

Two notable examples of modern mats are the pliant ones, made of pandanus leaves, found in southern Asia and Oceania and the tatami, which provide the unit of measurement of the surface area of Japanese dwellings. Just as basketry has been used for making containers and mats, so from ancient times to modern it has been used for making such pieces of furniture as cradles, beds, tables, and various kinds of seats and cabinets.

In addition to the use of basketry for skirts and loincloths (particularly common in Oceania), supple diagonal plaiting has even been used to make dresses (Madagascar). Plaited raincoats exist throughout eastern Asia as well as Portugal. Basketry most frequently is used for shoes (particularly sandals, some of which come close to covering the foot and are plaited in various materials), and, of course, for hats—the conical hat particularly common in eastern Asia, for example, and the skullcaps and brimmed hats found in Africa, the Americas, and much of Europe.

To protect head and body against weapons, thick, strong basketry has been used in the form of helmets (Africa, the Assam region in India, and Hawaii); armour (for example, armour of coconut palm fibre for protection against weapons made of sharks' teeth by the Micronesia inhabitants of the Gilbert Islands); and shields, for which basketry is eminently suitable because of its lightness. In addition to clothes themselves, there are numerous basketry accessories: small purses, combs, headdresses, necklaces, bracelets, and anklets. In West Africa there are even chains made of fine links and pendants plaited in a beautiful, bright yellow straw in imitation of gold jewelry. Many objects are plaited just for decoration or amusement such as ornaments like those used for Christmas trees or for harvest festivals and scale models and little animal or human figurines that sometimes serve as children's toys.

There is often no very clear distinction between accessories and ritual ornaments, as in the ephemeral headdresses made for initiation rites by the young Masa people in the Cameroon; dance accessories; ornaments for masks, such as the leaf masks that the Bobo of Upper Volta make with materials from the bush.

More clearly ritual in nature are the palms (woven into elaborate geometric shapes and liturgical symbols) carried in processions on Palm Sunday by Christians in various Mediterranean regions; some, like those from Elche in Spain, are over six feet (nearly two metres) high and take days to make. In Bali an infinite variety of plaiting techniques are involved in the preparation of ritual offerings, which is a permanent occupation for the women, a hundred of whom may work for a month or two preparing for certain great festivals.

Baskets as snares

Baskets are used throughout the world as snares and fish traps, which allow the catch to enter but not to leave. They are often used in conjunction with a corral (on land) or a weir (an enclosure set in the water), which are themselves made either of pliable nets or panels of basketry. In Africa as well as in eastern Asia a basketry object is used for fishing in shallow water; open at top and bottom, this object is deposited sharply on the bottom of shallow rivers or ponds, and, when a fish is trapped, it is retrieved by putting a hand in through the opening at the top.

Basketry is also used in harvesting foodstuffs; for example, in the form of winnowing trays (from whose French name, *van*, the French word for basketry, *vannerie*, is derived). One basket, found in the Sahel region south of the Sahara, is swung among wild grasses and in knocking against the stalks collects the grain.

Baskets are used as transport receptacles; they are made easier to carry by the addition of handles or straps depending on whether the basket is carried by hand, on a yoke, or on the back. The two-handled palm-leaf basket, common in North Africa and the Middle East, existed in ancient Mesopotamia; in Europe and eastern Asia, the one-handled basket, which comes in a variety of shapes, sizes, and types of plaiting, is common; in Africa, however, where burdens are generally carried on the head, there is no difference between baskets used for transporting goods and those used for storing.

Burden baskets are large, deep baskets in which heavy loads can be carried on the back; they are provided either with a headband that goes across the forehead (especially American Indian, southern Asia), or with two straps that go over the shoulders (especially in Southeast Asia and Indonesia). There are three fairly spectacular types of small basketry craft found in regions as far apart as Peru, Ireland, and Mesopotamia: the balsa (boats) of Lake Titicaca, made of reeds and sometimes fitted out with a sail also made of matting; the British coracle, the basketry framework of which is covered with a skin sewn onto the edge; and the gufa of the Tigris, which is round like the coracle and made of plaited reeds caulked with bitumen.

Origins and centres of development

Something about the prehistoric origins of basketry can be assumed from archaeological evidence. The evidence that does exist from Neolithic times onward has been preserved because of conditions of extreme dryness (Egypt, Peru, southern Spain) or extreme humidity (peat bogs in northern Europe, lake dwellings in Switzerland); because it had been buried in volcanic ash (Oregon); or because, like the mats at Jarmo, it left impressions in the mud or on a pottery base that had originally been molded onto a basketry foundation. More recently, when written and pictorial documentation is available, an activity as humble and banal as basketry is not systematically described but appears only by chance in narratives, inventories, or pictures in which basketry objects figure as accessories.

On the evidence available, researchers have concluded that the salient characteristics of basketry are the same today as they were before the 3rd millenium BC. Then, as now, there was a wide variety of types (and a wide distribution of most types): coiled basketry either spiral or sewed, including furcate and sewed braid (mainly in Europe and the Near East as far as the Indus valley); wattlework with twined threads (America, Europe, Egypt) and with woven threads (Jarmo, Peru, Egypt); and plaited construction with twilled weaving (Palestine, Europe).

To list the centres of production would almost be to list all human cultural groups. Some regions, however, stand out for the emphasis their inhabitants place on basketry or for the excellence of workmanship there.

American Indian basketry. In western North America the art of basketry has attained one of its highest peaks of perfection and has occupied a preeminent place in the equipment of all the groups who practice it. North American Indians are particularly noted for their twined and coiled work. The Chilkat and the Tlingit of the Pacific Northwest are known for the extreme delicacy of their twined basketry; the California Indians, for the excellence of their work with both types; and the Apache and the Hopi and other Pueblo Indians of the southwestern interior of the United States, for coiled basketry remarkable for its bold decoration and delicate technique.



Figure 150: Basketry mask, from the Sepik region of New Guinea. In the Musée de l'Homme, Paris.

By courtesy of the Musee de l'Homme, Paris

Central and South American basketry is similar in materials and plaiting processes. The notable difference lies in the finishes used, and in this the Guyana Indians of northeastern South America excel, using a technique of fine plaiting with a twill pattern.

Plaiting processes in Oceania

Oceanic basketry. Various plaiting processes have been highly developed in Oceania, not merely for making utilitarian articles but also for ceremonial items and items designed to enhance prestige, such as finely twined cloaks in New Zealand, statues in Polynesia, masks in New Guinea (Figure 150), and decorated shields in the Solomon Islands. In Oceania, as in southern Asia, there is a vegetal civilization, in which basketry predominates over such arts as metalwork and pottery. Particular mention should be made of the Senoi of the Malay Peninsula and of the Australian Aborigines, whose meagre equipment includes delicate basketry done by the women. The Senoi use various plaiting techniques, and the Australians use tight twining.

African basketry. Africa presents an almost infinite variety of basket types and uses. In such regions as Chad and Cameroon, basketry is in evidence everywhere—edging the roads, roofing the houses, decorating the people, and providing the greater part of domestic equipment. The delicate twill plaited baskets of the Congo region are notable for their clever patterning. In the central and eastern Sudanese zone the rich decorative effect of the sewed, coiled baskets is derived from the interplay of colours. People living in the lake area of the Great Rift Valley produce elegant coiled and twined basketry of restrained decoration and careful finishing.

East Asian basketry. People of the temperate zones of East Asia produce a variety of work. Bamboo occupies a particularly important place both in functional basketry equipment and in aesthetic objects (Japanese flower baskets, for example). The production of decorative objects is one feature that distinguishes East Asian basketry from the primarily utilitarian basketry of the Near East and Africa. Southeast Asia, together with Madagascar, are among the places known for their fine decorative plaiting techniques.

European basketry. In Europe almost the whole range of basketry techniques is used, chiefly in making utilitarian objects (receptacles for domestic and carrying purposes and household furniture) but also in making objects primarily for decorative use.

Modern basketry. Even in the modern industrial world, there seems to be a future for basketry. Because of its flexibility, lightness, permeability, and solidity, it will probably remain unsurpassed for some utilitarian ends; such articles, however, because they are entirely handmade, will gradually become luxury items. As a folk art, on the other hand, basketry needs no investment of money: the essential requirements remain a simple awl, nimble fingers, and patience.

(He.Ba.)

METALWORK

Man's first materials were stone, wood, bone, and earth. It was not until he had reached a later stage of evolution that he was able to extract and work with metals. This section deals with the objects that man has fashioned from copper, bronze and brass, gold and silver, iron, and lead: vessels, utensils, ceremonial and ritualistic objects, decorative objects, architectural ornamentation, personal ornament, sculpture, and weapons. It treats the processes and techniques and the stylistic characteristics and historical developments of metalworking.

General processes and techniques

Many of the technical processes in use today are essentially the same as those employed in ancient times. The early metalworker was familiar, for example, with hammering, embossing, chasing, inlaying, gilding, wiredrawing, and the application of niello, enamel, and gems.

Hammering and casting. All decorative metalwork was originally executed with the hammer. The several parts of each article were hammered out separately and then were put together by means of rivets, or they were pinned on a solid core (for soldering had not yet been invented). In addition, plates of hammered copper could be shaped into statues, the separate pieces being joined together with copper rivets. A life-size Egyptian statue of the pharaoh Pepi I in the Egyptian museum, Cairo, is an outstanding example of such work.

Lost-wax process

After about 2500 BC, the two standard methods of fabricating metal—hammering and casting—were developed side by side. The lost-wax, or cire perdue (casting with a wax mold), process was being employed in Egypt by about 2500 BC, the Egyptians probably having learned the technique from Sumerian craftsmen (see SCULPTURE, ART OF). Long after the method of casting statues in molds with cores had superseded the primitive and tedious rivetting process, the hammer continued as the main instrument for producing art works in precious metals. Everything attributable to Assyrian, Etruscan, and Greek goldsmiths was wrought by the hammer and the punch.

Embossing, or repoussé. Embossing (or repoussé) is the art of raising ornament in relief from the reverse side. The design is first drawn on the surface of the metal and the motifs outlined with a tracer, which transfers the essential parts of the drawing to the back of the plate. The plate is then embedded face down in an asphalt block and the portions to be raised are hammered down into the yielding asphalt. Next the plate is removed and re-embedded with the face uppermost. The hammering is continued, this time forcing the background of the design into the asphalt. By a series of these processes of hammering and re-embedding, followed finally by chasing, the metal attains its finished appearance. There are three essential types of tools—for tracing, for bossing, and for chasing—as well as a specialized tool, a snarling iron or spring bar, which is used to reach otherwise inaccessible areas. Ornament in relief is also produced by mechanical means. A thin, pliable sheet of metal may be pressed into molds, between dies, or over stamps. All of these methods have been known from antiquity.

Chasing. Chasing is accomplished with hammer and punches on the face of the metal. These punches are so shaped that they are capable of producing any effect—either in intaglio (incising beneath the surface of the metal) or in relief—that the metalworker may require. The design is traced on the surface, and the relief may be

obtained by beating down the adjacent areas to form the background. Such chased relief work sometimes simulates embossed work, but in the latter process the design is bossed up from the back. The detailed finish of embossed work is accomplished by chasing; the term is applied also to the touching up and finishing of cast work with hand-held punches.

Engraving. To engrave is to cut or incise a line. Engraving is always done with a cutting tool, generally by pressure from the hand. It detaches material in cutting. When pressure is applied with a hammer, the process is called carving.

Damascening and niello

Inlaying. The system of ornamentation known as damascening is Oriental in origin and was much practiced by the early goldsmiths of Damascus; hence the name. It is the art of encrusting gold wire (sometimes silver or copper) on the surface of iron, steel, or bronze. The surface upon which the pattern is to be traced is finely undercut with a sharp instrument. The gold thread is forced into the minute furrows of the cut surface by hammering and is securely held.

Niello is the process of inlaying engraved ornamental designs with niello, a silver sulfide or mixture of sulfides. The first authors to write on the preparation of niello and its application to silver were Eraclius and Theophilus, in or about the 12th century, and Benvenuto Cellini, during the 16th. According to each of these authors, niello is made by fusing together silver, copper, and lead and then mixing the molten alloy with sulfur. The black product (a mixture of the sulfides of silver, copper, and lead) is powdered; and after the engraved metal, usually silver, has been moistened with a flux (a substance used to promote fusion), some of the powder is spread on it and the metal strongly heated; the niello melts and runs into the engraved channels. The excess niello is removed by scraping until the filled channels are visible, and finally the surface is polished.

Champlevé and cloisonné

Enamelling. There are two methods of applying enamel to metal: champlevé, in which hollows made in the metal are filled with enamel; and cloisonné, in which strips of metal are applied to the metal surface, forming cells, which are then filled with enamel. (For a detailed discussion, see the section *Enamelwork* below.) (S.V.G.)

Gilding. Gilding is the art of decorating wood, metal, plaster, glass, or other objects with a covering or design of gold in leaf or powder form. The term also embraces the similar application of silver, palladium, aluminum, and copper alloys.

The earliest of historical peoples had masterly gilders, as evidenced by overlays of thin gold leaf on royal mummy cases and furniture of ancient Egypt. From early times, the Chinese ornamented wood, pottery, and textiles with beautiful designs in gold. The Greeks not only gilded wood, masonry, and marble sculpture but also fire-gilded metal by applying a gold amalgam to it and driving off the mercury with heat, leaving a coating of gold on the metal surface. From the Greeks, the Romans acquired the art that made their temples and palaces resplendent with brilliant gilding. Extant examples of ancient gilding reveal that the gold was applied to a ground prepared with chalk or marble dust and an animal size or glue.

Beating mint gold into leaves as thin as $\frac{1}{280,000}$ inch (0.00001 centimetre) is done largely by hand, though machines are utilized to some extent. After being cut to a standard $3\frac{3}{8}$ inches (9.84 centimetres) square, the leaves are packed between the tissue-paper leaves of small books, ready for the gilder's use.

The many substances to which the gilder can apply his art and the novel and beautiful effects he can produce may require special modifications and applications of his methods and materials. Certain basic procedures, however, are pertinent to all types of gilding. For example, the ground to be gilded must be carefully prepared by priming. Flat paints, lacquers, or sealing glues are used, according to the nature of the ground material. Metals subject to corrosion may be primed (and protected) by red lead or iron oxide paints. With pencil or chalk the gilder lays out his design on the ground after the ground has been prepared and is thoroughly dry. Patterns may also be laid down by forcing, or pouncing, powdered chalk or dry pigment through paper containing perforations made with pricking wheels mounted on swivels; the swivel arrangement permits the attainment of the most intricate of designs.

To create an adhesive surface to which the gold will be securely held, the area to be gilded is sized. The type of size used depends on the kind of surface to be gilded and on whether it is desirable for the size to dry quickly or slowly. When the size has dried enough so that it just adheres to the fingertips, it is ready to receive and retain the gold leaf or powder.

Application of gold leaf

Gold leaf may be rolled onto the sized surface from the tissue book. Generally, however, the gilder holds the book firmly in his left hand with the tissue folded back to expose as much leaf as is needed and detaches that amount with a pointed tool, such as a sharpened skewer. He then picks up the leaf segment with his gilder's tip, a brush of camel's hair set in a thin cardboard holder, and carefully transfers it to its place in his design. The leaf is held to the tip by static electricity, which the gilder generates by brushing the tip gently over his hair. For some gilding operations the gilder uses a cushion to hold his pieces of leaf. This is a rectangular piece of wood, about 9 by 6 inches (23 by 15 centimetres) in size, which is padded with flannel and covered with dressed calfskin; a parchment shield around one end protects the delicate leaf from disturbance by drafts of air. When the gilding is completed, the leaf-covered area should be pounced with a wad of soft cotton of surgical grade. Rubbing with cotton burnishes the gold to a high lustre. Application of a gilder's burnisher—that is, a highly polished agate stone set in a handle—also imparts a fine, high finish to the metal. Loose bits of gold, or skewings, may be removed from the finished work with a camel's hair brush.

Leaf gold may be powdered by being rubbed through a fine-mesh sieve. Powdered gold is so costly, however, that bronze powders have been substituted almost universally for the precious metal. When gold leaf is employed in the gilding of domes and the roofs of buildings, it is used in ribbon form. For finishing processes, such as burnishing and polishing, see SCULPTURE, THE ART OF. (E.L.Y.)

Western metalwork

COPPER

Copper alloys

The first nonprecious metal to be used by man was copper. But in the 4th millennium BC, Eastern craftsmen discovered that copper alloys using tin or zinc were both more durable and easier to work with, with the result that from then on the use of unalloyed copper declined sharply. Artists and craftsmen working in the West also discovered this, which is why pure copper work was relatively rare.

Pure copper is a reddish colour and has a metallic glow. When it is exposed to damp, it becomes coated with green basic copper carbonate (incorrectly known as verdigris). This patina is a drawback if copper is to be used for functional objects, for the oxide is poisonous to man. This means that utensils that come into contact with food must be lined with tin.

As copper is a relatively soft metal, it is sensitive to such influences as stress and impact. But unlike bronze it is malleable and can be hammered and chased in much the same way as silver. The surface of copper can be successfully gilded, and its reddish colouring makes the gilding seem even brighter. Because of these properties, copper was sometimes able to compete somewhat with silver.

Pure copper is not particularly good for casting, as it can easily become blistered when the gases escape. The surface of sheet copper can be engraved, however, and this technique was often used for decorating purely ornamental objects. In copperplate etching, engraving became the basis of printing. Enamel is often applied to copper, using both the champlevé and cloisonné techniques. Sheet copper was also used as a base for painted enamel. (H.-U.H.)

Antiquity. *Mesopotamia.* In the museum at Baghdad, in the British Museum, and in the University of Pennsylvania at Philadelphia are finely executed objects in beaten copper from the royal graves at Ur (modern Tall

Copper relief at al-'Ubaid

al-Muqayyar) in ancient Sumer. Outstanding is a copper relief that decorated the front of the temple at al-'Ubaid. This remarkable decoration represents an eagle with a lion's head, holding two stags by their tails. The stags' antlers—also made of wrought copper—were developed in high relief and were soldered into their sockets with lead. This relief illustrates the high level of art and technical skill attained by the Sumerians in the days of the 1st dynasty of Ur (*c.* 2650–2500 BC). In the Metropolitan Museum of Art, New York City, is a Sumerian bull's head of copper, probably an ornamental feature on a lyre, which is contemporary with the Ur finds.

The malleability of unalloyed copper, which renders it too soft for weapons, is peculiarly valuable in the formation of vessels of every variety of form; and it has been put to this use in almost every age. Copper domestic vessels were regularly made in Sumer during the 4th millennium BC and in Egypt a little later.

Egypt. From whatever source Egypt may have obtained its metalworking processes, Egyptian work at a remote period possesses an excellence that, in some respects, has never been surpassed. Throughout Egyptian history, the same smiths who worked in the precious metals worked also in copper and bronze.

Nearly every fashionable Egyptian, man or woman, possessed a hand mirror of polished copper, bronze, or silver. Copper pitchers and basins for hand washing at meals were placed in the tombs. An unusual example in the Metropolitan Museum of Art is plated with antimony to imitate silver, which was very rare in the Old Kingdom (*c.* 2686–*c.* 2160 BC). The basins and the bodies of the ewers were hammered from single sheets of copper. The spouts of the ewers were cast in molds and attached to the bodies by means of copper rivets or were simply inserted in place and crimped to the bodies by cold hammering.

(S.V.G.)

Middle Ages. *Europe.* The first well-designed copper objects to survive in the West date from about the middle of the Carolingian period, the 8th century AD. Who made them is not known, but one can assume that in the early Middle Ages they were mainly the work of monks. Indeed, the earliest copper and copper-gilt pieces are exclusively liturgical implements.

Decrees issued by the church synods held in the 8th and 9th centuries invariably expressly prohibited the use of copper and bronze for consecrated chalices, but in fact a few copper-gilt chalices like the "Tassilo Chalice" (Kremsmünster Abbey, Austria) have survived (Figure 151). The care and artistry with which they were worked and their rich engraved and niello decoration show that they were valued as highly as altar vessels made of precious metals.



Figure 151: "Tassilo Chalice," copper gilt with silver and niello, c. 780. In the Kremsmünster Abbey, Austria. Height 25 cm.

Bildarchiv Foto Marburg

Copper-gilt religious objects

From the 12th century onward, but particularly in the 13th and 14th centuries, copper-gilt chalices were relatively common, especially in Italy, where they were virtually mass-produced. Reliquaries, portable altars, shrines, and processional crosses dating from the Ottonian and Romanesque periods are also very frequently made of gilded copper and are generally decorated with enamel, niello work, or engraving or set with precious stones. One group of copper-gilt reliquaries, dating from the 12th century and after, takes the form of the head, or head and shoulders, of a saint. Others are in the shape of various parts of the body, such as an arm or a foot. These were also made in silver and in cast bronze. Ciboria (covered vessels for holding the wafers of the Eucharist), monstrances (receptacles for the Host), incense vessels, and other liturgical implements were also made in copper gilt, as well as in bronze and silver. Some of these copper-gilt implements were made as late as the Baroque period. (H.-U.H.)

Islām. The most magnificent example of Muslim enamel work in existence is a copper plate in the Tiroler Landes museum Ferdinandeum at Innsbruck, Austria, decorated in polychrome enamel, with figure subjects, birds and animals within medallions separated by palm trees and dancers (first half of the 12th century). The Mesopotamian, or Mosul, style, which flourished from the early part of the 13th century, is characterized by a predominant use of figures of men and animals and by the lavish use of silver inlay. The most famous example of figured Mosul work in Europe is the so-called Baptistery of St. Louis in the Louvre. This splendid bowl, which belongs in style to the Mosul work of the 13th century, measures five feet (150 centimetres) in circumference and is covered with figures richly inlaid with silver, so that little of the copper is visible. It is signed by the artist. (S.V.G.)

Renaissance to modern. In the second half of the 16th century, copper gilt began to be used less and less often for liturgical implements because silver had become cheaper and was therefore preferred.

In the late 16th century, Italian smiths used copper for water beakers and water jugs, decorating the surfaces with chased ornaments, whereas the rest of Europe used brass.

High-quality copper objects dating from the 17th and 18th centuries were sometimes designed and worked in the same way as the silver of the period. Most were probably trial pieces made for the guild rank of journeyman or master by silversmiths who were too poor to supply objects in precious metal. Some may have been used as workshop models or given to clients as specimen pieces.

Another type of copper vessel, known as a "Herrengrund cup," is purely ornamental and resembles the showpieces made in the 16th and 17th centuries. These mugs are made of copper that was extracted by a process known as cementation, in which water containing copper forms a deposit on iron. Production was limited to three places in the county of Sohl in Hungary. In those days the process seemed mysterious to many people; many of the inscriptions on "Herrengrund cups" refer to this mystery. The design of the beakers is modelled closely on that of silver vessels produced in southern Germany, Bohemia, and Silesia. The best examples are chased, engraved, or gilded or, more rarely, enamelled or set with precious stones. Many of them are decorated with mining scenes peopled with little figures. Most were made in the 17th century; a decline set in in the 18th century, though individual pieces continued to be made until the Empire period.

In the 17th and 18th centuries, copper enjoyed a period of relative prosperity in middle class households on the continent of Europe. For example, copper bread bins lined with tin were used; they were often richly decorated with chased motifs or brass fittings. There were also sumptuous wine coolers, cake and pudding molds, bowls, buckets, jugs, jars, screw-top flasks, sausage pans, and many other items, all polished until they shone and thus used as kitchen decorations as well as utility items.

In 18th-century Holland, jugs for tea and coffee were made in copper with a dark-brown patina and with various parts, such as the handle and the knob, in brass gilt. The sides were chased with interlaced foliage and other Rococo decorative motifs.

Sheffield plate

Copper was also the main metal used for Sheffield plate, which has a silvered surface. In 1742 Thomas Bolsover invented a method of fusing copper and silver together so that the result was highly durable, and he produced this type of silver-plated ware on a large scale. Although 18th-century England was a relatively wealthy society and solid silver utensils of all kinds were used fairly widely, the middle classes, who were not all that well off, liked to buy these implements that looked like silver yet cost only a third of the price. The makers of Sheffield plate therefore adopted the designs used for English silverware at that date, and their work was often as courtly and elegant as that of the silversmiths (Figure 152).

Crown Copyright. Victoria and Albert Museum, London

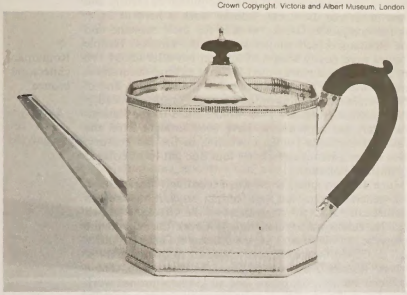


Figure 152: Sheffield plate teapot, English, late 18th century. In the Victoria and Albert Museum, London. Height 16.5 cm.

Copper ware was no longer important in the 19th century, though it was occasionally used for pieces designed to follow earlier styles or for copies of historical pieces. The method now used was electroplating, which is a purely technical process and has nothing to do with craftsmanship.

Toward the end of the 19th century, attempts were made to create a new and individual style for copper; and there were occasional signs that its inherent properties were understood and used to full effect. But there was no renaissance in the true sense of the word.

BRONZE AND BRASS

Bronze is an alloy of copper and tin. In the period of classical antiquity it had a low tin content, generally containing less than 10 percent, because tin was less common and therefore difficult to obtain. Like bronze, brass is an alloy, this time of copper plus zinc.

It is often very difficult to distinguish between bronze and brass merely by their appearance. The colour of the different alloys ranges over various shades from gold to a reddish tinge, to silvery, greenish, and yellowish shades, according to the proportions of the basic constituents. The patina on both alloys ranges from dark brown to a dark greenish tinge, particularly in the earliest pieces. Since it is often difficult to differentiate between bronze and brass with the naked eye and since metalworkers and metal casters of previous centuries did not make an express distinction between them, they will be considered together here. From a very early date bronze was used mainly for casting. Because it is so brittle, it has only rarely been hammered or chased; brass or copper were preferred for such work because they are more malleable. Down to the Middle Ages, bronze was cast by the cire perdue, or lost-wax, method. By this process, the mold can be used only once. This method of casting is the most exclusive, not only because it is the most expensive but also because it produces the finest work from the aesthetic point of view. Later, the casting process used models made up of a number of different pieces that could be taken apart and therefore re-used. These were generally made of wood and could be pressed down into a sand mold so that the shape of the object being cast emerged as a hollow. The hollow was then filled with molten bronze, which was poured in through casting ducts. When the resulting piece had been removed from the sand mold, the surface was smoothed over and the casting seams removed. The wooden model could then be used again to make as many copies as were required, which meant that economical production was possible. Brass was cast by the same methods but over and above this a process of hammering and chasing was used to fashion sheet brass. Brass platters were often decorated with relief work ornament, which was embossed from the reverse side by means of a type of die. The brass worker could also create an ornamental frieze made up of small motifs by using a series of punches made of iron. The surface of bronze or brass objects was also occasionally decorated with engraving.

Bronze casting

(H.-U.H.)

Antiquity. *Mesopotamia.* In the Metropolitan Museum of Art is the bronze sword of King Adad-nirari I, a unique example from the palace of one of the early kings of the period (14th–13th century BC) during which Assyria first began to play a prominent part in Mesopotamian history. A magnificent example of Assyrian bronze embossed work is to be seen in the gates of Shalmaneser III (858–824 BC), erected to commemorate that king's campaigns. The gates were made of wood; and the bronze bands, embossed with a wealth of figures in relief, are only about $^{1}/_{16}$ inch (1.6 millimetres) thick. The bands were obviously intended for decoration, not to strengthen the gates against attack.

Iran. The Persian bronze industry was also influenced by Mesopotamia. Luristan, near the western border of Persia (Iran), is the source of many bronzes that have been dated from 1500 to 500 BC and include chariot or harness fittings, rein rings, elaborate horse bits, and various decorative rings, as well as weapons, personal ornaments, different types of cult objects, and a number of household vessels. Many of these objects show a decided originality in the development of the animal style (Figure 153).



Figure 153: Cast bronze finial from Luristan, 9th–8th century BC. In the Denver Art Museum, Colorado. Height 16 cm.

By courtesy of the Denver Art Museum, Colorado

Egypt. The bronzes that have survived are mainly votive statues placed in the temples from the Saite to the Ptolemaic period (305–30 BC), and amuletic bronzes that were buried with the dead. In its simplest form the decoration consisted of lines, representing details of clothing, ornaments, and the like, cut in the bronze with engraving tools, sometimes also combined with gilding. A fine example of inlay work of the 22nd dynasty (945–*c.* 730 BC) is a bronze menat damascened with gold wire (Metropolitan Museum of Art).

Crete. A sword, found in the palace of Mallia and dated

Dagger blades

to the Middle Minoan period (2000–1600 BC), is an example of the extraordinary skill of the Cretan metalworker in casting bronze. The hilt of the sword is of gold-plated ivory and crystal. A dagger blade found in the Lasithi plain, dating about 1800 BC (Metropolitan Museum of Art), is the earliest known predecessor of ornamented dagger blades from Mycenae. It is engraved with two spirited scenes: a fight between two bulls and a man spearing a boar. Somewhat later (*c.* 1400 BC) are a series of splendid blades from mainland Greece, which must be attributed to Cretan craftsmen, with ornament in relief, incised, or inlaid with varicoloured metals, gold, silver, and niello. The most elaborate inlays—pictures of men hunting lions and of cats hunting birds—are on daggers from the shaft graves of Mycenae, Nilotic scenes showing Egyptian influence. The bronze was oxidized to a blackish-brown tint; the gold inlays were hammered in and polished and the details then engraved on them. The gold was in two colours, a deeper red being obtained by an admixture of copper; and there was a sparing use of neillo.

Greece. The Greeks, who learned much about metalwork from the Egyptians, excelled in hammering, casting, embossing, chasing, engraving, soldering, and metal intaglio. Among the ancients, the great emphasis of technology was on aesthetic expression, not on practical utilization. Greek coin dies rank with the finest work of this kind that the world has ever seen. Pottery and bronze hammer-and-cast work were important crafts of ancient Greece. Vases of terra-cotta were often designed to resemble those of bronze, and both kinds were widely used in antiquity. Unlike terra-cotta, which is breakable but otherwise practically indestructible, bronze is subject to corrosion; and a surviving Greek bronze vase in good condition is therefore something of a rarity. The body of the vase, which was hammered out of a sheet of malleable bronze, was usually left plain; the handles, feet, and overhanging lip, which were cast, were decorated. The applied elements were rivetted or soldered.

It was in the time of Lysippus, the distinguished sculptor who flourished about 330 BC, that the fine Greek beaten work for decoration of armour, vases, and objects of domestic use reached its perfection. It was executed by a hammer worked from behind, the outlines being afterward emphasized by chisel or punch; or metal plate was beaten into a mold formed by carving the subject in intaglio upon a resisting material. The embossed shoulder straps of a cuirass, called the "Bronzes of Siris" (4th century BC; British Museum, London), are in exceedingly high relief and are beaten into form with wonderful skill with the hammer. The relief depicts the combat between the Greeks and the Amazons.

Statuettes and monumental sculpture

Greek bronze statuettes—originally dedicatory offerings in shrines, ornamental figures on utensils, or decorative works of art—have survived in large numbers. They were usually cast solid, rarely hollow. Sometimes even large statuettes were cast solid. (The advantage of solid casting is that the mold can be used repeatedly, whereas in the hollow-casting process the mold is destroyed.) Greek bronzes were originally golden and bright, and they were often decorated with silver or niello for colour contrast. Bronze statuary hardly existed before the introduction of hollow casting, about the middle of the 6th century BC, after which bronze became the most important medium of monumental sculpture; its strength and lightness admitted poses that could not be reproduced in stone.

Etruria. The Etruscans used bronze for cast and beaten work; and although few large works remain, the museums of Europe display a marvellous variety of admirably formed small bronzes. A masterpiece of bronze Etruscan sculpture is the "Chimera" (a mythological beast with a goat's body, a lion's head, and a serpent's tail) from Arezzo, a 5th-century BC ex-voto from a sacred building, found in 1553 and partly restored by Benvenuto Cellini (Museo Archeologico di Firenze). Etruscan bronze workers produced, often for export, votive statuettes, vessels, furniture, helmets, swords, lamps, candelabra, mirrors, and even chariots. An Etruscan chariot of *c.* 600 BC in the Metropolitan Museum of Art has a body and wheels of wood, sheathing of bronze, and tires of iron, the high front embossed with archaic figures of considerable grace. The Etruscans inlaid bronze with silver and gold in a manner that proves that their skill in this mode of enrichment equalled that of the Greeks and Romans. Many delicately engraved bronze objects were made in the Latin town of Praeneste (modern Palestrina), which possessed a highly developed bronze-working industry. From Praeneste came a remarkable cylindrical container of the late 4th century BC, now in the Villa Giulia, Rome; its richly engraved surface provides a good example of the perfection of ancient drawing.

Rome. Etruscan cities, like those of Greece, were crowded with bronze statues of gods and heroes; and Rome derived its best adornment from the pillage of Etruria and then of Greece. Distinctly Roman work is hard to trace, as the conquered Greeks worked for their masters, and the Romans copied wholesale from the Greeks. Temple statues were nearly always of bronze, but after about 190 BC the metal was chiefly used for architectural decorations and portraiture. The bronze doors of the Pantheon and of the Temple of Romulus in the Roman Forum still occupy their original positions. Two bronze doors in the Lateran Baptistery are supposed to have been brought from the Baths of Caracalla by Pope Hilarius in the 5th century. Also in the Lateran church are four fine gilt-bronze fluted Corinthian columns.

Roman architectural bronze

Much Roman small work was exceedingly fine, though it is generally conceded that Roman productions are less aesthetically attractive than those of the Greeks. Pompeii and Herculaneum were essentially Greek towns, and the many beautiful bronzes in the Museo e Gallerie Nazionali di Capodimonte, Naples, collected from the ruins of private houses there, are of Greek workmanship. These included statuettes, mirrors, and all kinds of bronze work useful in a house. Many of these pieces were originally attached to pieces of furniture.

During the closing years of the republic, brass, produced by what came later to be known as the calamine (zinc-carbonate) method, became an important material for the first time. Its various uses included parade armour, as may be seen in a Roman embossed brass helmet in the Castle Museum, Norwich, England.

Teutonic tribes. The Teutonic tribes who conquered and divided the Roman Empire were little versed in the monumental arts and unskilled in figure representation; but in metalworking, in the making of weapons and other utilitarian objects, and in the delicate ornament of the goldsmith's art they excelled. They were among the earliest in Western Europe to develop the use of enamel decoration on bronze in the champlevé technique.

Middle Ages: Byzantine Empire. Syria, Egypt, and Anatolia were first the teachers and then the rivals of Constantinople (Istanbul). The fusion of antique and Eastern elements resulted in the Byzantine style, the great period of which dates from the 9th to the end of the 12th century. The extensive use of embossed work, with filigree, cabochon gems, and small plaques of enamel, may be seen in both the East and the West during the early Middle Ages. The most conspicuous examples of large Byzantine metalwork are bronze church doors inlaid with silver. Many objects are still preserved in various European treasuries, which were enriched by the spoils of the sack of Constantinople in 1204. Venice, in the Treasury of St. Mark's, has an unrivalled series of Byzantine chalices, bookbindings, and other treasures of metalwork; but it is in Kiev, Moscow, and Leningrad that broadly representative series of all the categories of Byzantine artistic production may be found.

The art of bronze casting had been preserved in the Byzantine Empire. The first bronze doors to be made after the art had died out in Rome were those for Hagia Sophia at Constantinople, which bear the date 838; the panels, with monograms and other ornament damascened in silver, are framed in borders cast in relief and enriched with bosses and scrolls, the whole in an admirable style. Two sets of doors in St. Mark's, Venice, of Greek workmanship and considerable but uncertain antiquity, are supposed by some to have been removed from St. Mark's at Alexandria. Next in date among surviving doors of

Bronze doors of Hagia Sophia

Byzantine workmanship is a series ordered by the Pantaleone family (about 1066–87) and destined for cities in southern Italy—Amalfi, Trani, Salerno, Canosa di Puglia, and Monte Sant'Angelo. (S.V.G.)

Middle Ages: Islām. Animals in the Sāsānian style—lions, dragons, sphinxes, peacocks, doves, cocks, and the like—were cast in bronze in three dimensions and served, like their ceramic counterparts, as basins, braziers, and so on. They were particularly sought after in the later Abbāsid, Fātimid, and Seljuq periods, and from Egypt they became prototypes of similar European forms. It was the Seljuqs, apparently, who introduced a round bronze mirror, the reverse of which shows in low relief two sphinxes face to face, surrounded by a twined pattern, or two friezes with the astrological symbols of the seven chief heavenly bodies (Sun, Moon, and the five nearest planets) and the 12 signs of the zodiac, surrounded by a band of script; this goes back ultimately to Chinese origins.

Early vessels, such as mugs, were ornamented with animals in low relief, but engraving quickly supplanted this. Under the later Seljuqs (particularly the Artuqid atabegs of Mosul) and the Mamlūks, engraving became almost the only form of decoration, but only to serve as a basis for the yet richer technique of inlaying, or damascening: small silver plates and wires, themselves delicately engraved, were hammered into the ribs and surfaces, which were hollowed out and undercut at the edges.

Engraving and inlaying

In place of this, in an Artuqid bowl in the provincial museum at Innsbruck the spaces are filled in with cellular enamel. This was a method of evading the prohibition of precious metals, just as gold lustre was in pottery. The ornament consisted of friezes and medallions in lattice work and arabesque work, the interstices being filled with figures of warriors, hunters, musicians, animals, and astrological symbols. These were superseded later by Mamlūk coats of arms and inscriptions. In the 15th century the technique was imported from Syria to Venice, where productions of the same kind, *alla damaschina* or *all'azzimina*, were made right into the 16th century by Islāmic masters and were in great demand. In the East the process is still common, but both technically and artistically it has decayed.

In the 15th century there was a renaissance of pure metal engraving, but the design—inscriptions and arabesques in the Timūrid and Safavid styles—was not cut into the material but left free in the manner of a relief, the background being etched in black. Decoration was applied to bowls, basins, mugs, vases, mortars, braziers, warming pans, candlesticks, smoking utensils, inkstands, jewel cases, Qur'ān holders, and mosque lamps. These are generally in the simplest possible forms—spherical, cylindrical, prismatic; the subjects include motifs of vegetation and animal life, the former mainly in the necks and feet of vessels, the latter for handles and ears, feet, and sometimes small spouts. (H.Go.)

Revival of the art of bronze casting

Europe from the Middle Ages. After several centuries of artistic decline, the art of bronze casting was revived in *c.* 800 by Charlemagne, who had monumental bronze portals made for the Palatine Chapel in his residence in Aachen, with bronze grilles placed inside it. The artists, who probably came from Lombardy, followed the styles of classical antiquity.

For many centuries the Christian Church remained the bronze caster's chief patron. Like the stonemasons, who also were heavily patronized by the church, they joined together to form associations, or foundries. These casting foundries hired themselves out to the large ecclesiastical building sites. They cast bells—almost every church had at least one bell—and monumental doors decorated with relief work; for instance, doors for Mainz (*c.* 1000) and Hildesheim (1015) cathedrals, for the cathedrals at Gneissen and Augsburg (11th century), and for St. Zeno Maggiore in Verona (12th century). They also made large fonts, the most famous being the one made by Renier de Huy in 1107–18 for the church of Notre Dame aux Fonts in Liège (now in the church of St. Barthélemy in Liège). The Dinant workshops, which formed the main centre for bronze casting in the Meuse district in the Middle Ages, specialized in what are known as "eagle lecterns." These are book stands with ornamental pedestals, with the panel supporting the enormous missals taking the form of the outspread wings of an eagle, a griffin, or a pelican. The earliest documented eagle lectern was made in 965, but the earliest example to have survived dates from 1372. It was made by Jean Joses of Dinant for the Church of Our Lady at Tongeren (Tongres), near Liège.

Records show that from the 11th to the 15th century there were more than 50 monumental seven-branched candlesticks (menorah) in various churches in Germany, England, France, Bohemia, and Italy, though only a few of these have survived. Documents relating to the Carolingian period speak of monumental bronze crucifixes and statues of the Virgin and of the saints, though the earliest surviving statues date from the 11th century; the crucifix in the abbey church at Werden, for example, dates from *c.* 1060 and was probably cast in a foundry in Lower Saxony.

Among the most outstanding examples of figurative bronze sculpture dating from the Romanesque period are a group of reliquaries designed in the shape of heads or heads and shoulders or occasionally arms, hands, or feet, according to the type of relics they contain. They were made in Lower Saxony or in France.

A few large chandeliers have survived from the 11th and 12th centuries, representing a sort of halfway stage between sculpture and functional objects. A far larger number are known to have existed from documents and contemporary accounts, but these have disappeared over the centuries. Examples from Germany, the southern half of the Low Countries, and France have survived or are documented. Romanesque chandeliers are always designed in the form of a crown. Candleholders, with architectonic structures and figures placed in between them, project from the crown.

Besides the monumental bronzes that have survived from the 8th to the 12th century, there are also a number of smaller pieces, such as processional crosses, altar crucifixes, chests, reliquaries, and similar articles. Another group of liturgical objects consists of candlesticks used to adorn altars. Their design often shows a wealth of invention, and they are decorated in the most sumptuous fashion. There was yet another group of candlesticks, which were secular in nature, that embodied the ideal of chivalry. They are cast in the shape of human figures: an armed warrior on horseback bearing a candleholder with a spike on which the candle is placed; a kneeling page in court dress holding a candle socket in his outstretched hands; or Samson perched on the lion's back, brandishing a candleholder. These candlestick figures are rare and precious examples of courtly life in the Romanesque period in Germany, France, England, and Scandinavia. Even at that time they were thought of as rare, deluxe articles within the reach of only a few privileged people.

Candlesticks

By courtesy of the Metropolitan Museum of Art, New York, The Cloisters collection, purchase, 1947

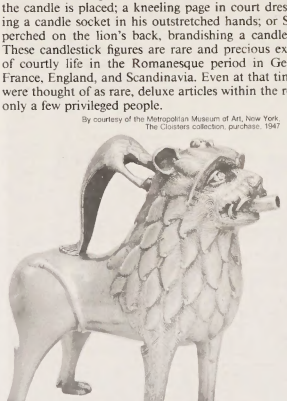


Figure 154: Bronze aquamanile in the shape of a lion, German, 13th century. In the Metropolitan Museum of Art, New York. Height 26.7 cm.

Toward the end of the Romanesque period a simpler type of candlestick appeared, mainly intended for religious purposes, though they were found in private homes as well. They are circular, with a round base, a slender column-like shaft, and a large grease pan with a spike for the candle. This design exercised a strong influence throughout the Gothic period and right down to the Baroque period, though it varied considerably over the years according to the styles then prevailing.

Aquamaniles and basins

Some of the finest bronze articles of the High Middle Ages were modelled on Oriental pieces brought back from the Holy Land by the crusaders. They are known as aquamaniles, a type of ewer used for pouring water for washing one's hands. Made by bronze casters in France, Germany, England, and Scandinavia, they are usually in the shape of lions—symbols of valour, pride, physical strength, and power (Figure 154). Also common are those shaped like knights in armour, with a wealth of courtly detail that was obviously popular. A few aquamaniles are in the shape of winged dragons, doves, cockerels, centaurs, or sirens; but such designs are rare. Christian themes, too, played a part, some examples depicting Samson overcoming the lion with his knee planted on its back. The golden age of these vessels was the 12th, 13th, and 14th centuries. The end of the age of chivalry also saw a decline in such work, for the emergent bourgeoisie found other ways of marking the ceremony of hand washing.

Basins were also needed for washing one's hands; they are often mentioned in medieval documents, where they are referred to as *bacina, pelves,* or *pelvicula.* The majority of these bowls—which date from the 12th and 13th centuries—have been found in the cultural area that extends from the Baltic down to the Lower Rhine district and across to England. Because this area was once dominated by the Hanseatic League (a commercial association of free towns), the basins are known as Hanseatic bowls. They are round, some being more convex than others; and the inside is engraved with scenes from classical mythology, with themes from the Old and New Testaments and the legends of the saints, or with allegorical figures personifying the virtues and the vices, the liberal arts, the seasons, and so on. Hanseatic bowls were probably made in the bronze-casting centres where candlesticks and aquamaniles (and indeed all medieval cast bronze) were made: in the Meuse district and Lorraine, in Lower Saxony and the Harz Mountains, and also in England. The decoration on these bowls may have been added elsewhere.

Alinari—Art Resource



Figure 155: Bronze doors from the north side of the Baptistery in Florence, by Lorenzo Ghiberti, *c.* 1403–24.

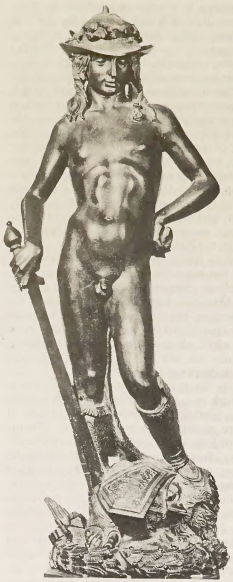


Figure 156: "David," bronze sculpture by Donatello, *c.* 1430–35. In the Bargello, Florence. Height 1.58 m.

Anderson—Alinari from Art Resource

In the Romanesque period and later, in the Gothic period, the churches and their patrons were still the bronze caster's main clients, ordering both functional objects and decorative pieces. Bronze fonts were relatively common in the 14th and 15th centuries, particularly in churches in northern Germany. Another common item, which was made mainly in England and in the Netherlands, was a large brass tombstone decorated with engraving. Other objects included door fittings, candlesticks, candelabra, chandeliers, pulpits, and sculptured tombs portraying the deceased.

Italy. Until the 12th century in Italy the art of bronze casting had been virtually neglected since the period of classical antiquity, when it had been a flourishing industry. A few churches in Italy have bronze doors inlaid with Byzantine niello work made by Byzantine craftsmen in the 11th and 12th centuries. The same technique was used by Bohemond I of Antioch for a bronze door at Canosa (1111) and by Oderisius of Benevento when casting a pair of doors for Troia Cathedral in 1119 and 1127. In the

Byzantine influence in Italy

second half of the 12th century, however, Barisano da Trani made relief door panels for churches in Astrano, in Ravello (a town near Amalfi), and in Monreale. Bronze relief doors were also made in the 12th century for S. Paolo fuori le mura in Rome and for churches in northern Italy (S. Zeno Maggiore in Verona; St. Mark's in Venice) and Tuscany (Pisa and Monreale, by Bonanno of Pisa) and in the 13th century for the Baptistery in Florence, by Andrea Pisano.

Lorenzo Ghiberti's doors for the Baptistery in Florence, made in 1403–24 and 1425–52 (Figure 155), marked the beginning of a golden age of bronze casting in Florence that lasted throughout the Renaissance and right down to the Baroque era. Whereas bronze sculpture had been relatively rare before the 15th century, many Italian artists of the Renaissance now designed cast bronze statues, statuettes, reliefs, and various objects in the shape of human figures. Among the sculptors who worked in full-scale bronzes were Lorenzo Ghiberti, Donatello (Figure 156), Andrea del Verrocchio, Antonio Pollaiuolo, and Lucca della Robbia. Besides large-scale cast-bronze work there were also small figures, statuettes, busts, plaques, and functional objects such as candelabra, mortars, candlesticks, and inkwells. Dating from the middle of the 15th century onward, they are characterized by rich figural and ornamental design. Their style influenced work produced in northern Europe, particularly in the 16th century.

In the first half of the 16th century, bronze casting declined somewhat in Italy, though it found a new lease on life in the middle of the century and, indeed, became even more important than before. Benvenuto Cellini and Giovanni da Bologna are two of the most famous artists of this period. Cellini designed a number of statues, one of the best known being his "Perseus" in the Loggia dei Lanzi in Florence, as well as portrait busts, reliefs, and smaller articles in bronze. Giovanna da Bologna, a Fleming by birth, was active in Rome and Florence, where he made fountains, equestrian monuments, allegorical figures, crucifixes, statuettes, groups of figures, animals, and many other objects. He founded a school of sculptors who were influenced by his work for many years. Many other bronze sculptors were active in the 16th and 17th centuries, notably in Venice, which was a particularly fruitful area for bronze casting, and at a school in Padua led by Andrea Riccio (Briosco). Italian bronze casters worked abroad as well as in their homeland, working on commission for foreign potentates, mainly in France and England.

In the 16th century, beautifully made bronze pieces, which were very much more than functional objects, played an important part in the art of the bronze caster. For instance, sumptuous mortars were designed and made by artists whose names have been handed down to posterity, such as Cavadini, Lenotti, Juliano da Navi, Alessandro Leopardi, Antonio Viteni, and Crescimbeni da Perugia. Elaborate brass dishes were made in Venice, under the influence of Eastern art (to which Venice had always been very receptive); indeed, the first people to produce these large dishes with engraved motifs were Islāmic artists who had settled in the town, though the local artists soon adopted both their style and their technique.

Germany and the Low Countries. Unlike their Italian counterparts, 15th-century bronze artists in Germany and the Low Countries were still under the spell of Gothic art, and ecclesiastical implements predominated.

The Dinant workshops, in the Meuse district, continued to dominate production until well past the middle of the 15th century, just as they had since the days of Charlemagne. But when Philip III the Good, duke of Burgundy, laid siege to the town in 1466, then took it by storm and eventually completely destroyed it, the bronze casters who survived moved elsewhere, settling mainly in the Low Countries. As a result, from that date onward the trade enjoyed a sudden upsurge in Brussels and Namur, in Tournai and Bruges (Flemish Brugges), in Malines (Flemish Mechelen), Louvain (Flemish Leuven), and Middelburg. There was another centre of the bronze trade in Lower Saxony, since the mines in the Harz Mountains produced a generous supply of copper and calamine. The chief bronze-working towns in this area were Hildesheim, Goslar, and Minden. In the 16th century, a period when trade and commerce were developing very rapidly in Germany, the bronze-casting trade was no longer compelled to function close to the place where the raw material was extracted. Thus, Nürnberg, at this time the most powerful and lively town in Germany, not only traded in copper, bronze, and brass but also soon allowed its bronze casters and metalworkers to develop a flourishing industry. Brass articles from Nürnberg became famous throughout the world.

Nürnberg brass work

The earliest documented brass workers were those known as "basin-beaters" (*Beckenschläger*), who were first referred to as such in 1373. They made bowls and dishes with various types of relief decoration on the bottom. In the late Gothic period, religious themes were very popular for this decoration and were more common than secular images (Figure 157). During the Renaissance, beginning

Crown Copyright Victoria and Albert Museum, London



Figure 157: Brass dish with embossed Annunciation scene. German c. 1500. In the Victoria and Albert Museum, London. Diameter 45 cm.

in about 1520, the design changed; instead of deep bowls there were large, flat dishes with decoration that consists of purely ornamental motifs or friezes as well as scenes and figures. The decoration includes the typically Gothic "fishbladder" design and also interlaced motifs and bands of lettering. The trade of the basin beaters continued to flourish in Nürnberg down to about 1550, when a decline set in, culminating in its eventual collapse just before the Thirty Years' War in 1618. The reason for this decline may have been the emergence of what is known as display pewter (see below *Pewter*), which, from about 1570 onward, swept the wealthy bourgeoisie market.

Until the Gothic era, bronze chandeliers were made solely for the churches; it was not until the 15th century that people began to consider lighting their homes by means of a central source of light hanging from the ceiling. In the Low Countries, one of the centres of the art of bronze casting, a type of chandelier was developed at this time that remained standard for many years. It is a type of hoop with a shaft, made up of a molded vertical centrepiece and a series of curving branches bearing drip trays and spikes. The arms, or branches, are decorated with tracery, foliage scrolls, and other motifs characteristic of the late Gothic style (Figure 158). In the middle of the 16th century, the central shaft took on the shape of a spherical baluster, with a large sphere jutting out just below the point where the curving arms branch off. This design continued to predominate in the Baroque period and is found as late as the 18th century. Because chandeliers of this type were most common in the Low Countries, one can assume that they originated there and were produced in large numbers and that they spread to England and Germany. Another centre was in Poland, presumably because brass founders had moved there from Nürnberg.

Besides these chandeliers—which until the 19th century

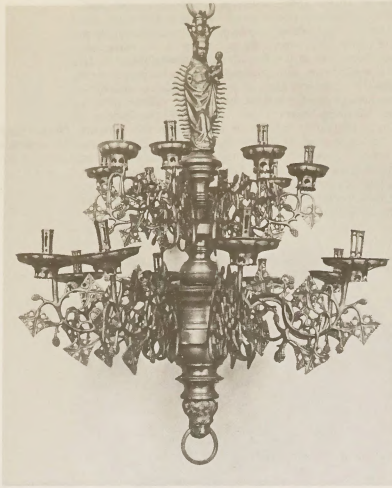


Figure 158: Bronze chandelier, Dutch, 15th century. In the Rijksmuseum, Amsterdam. Height 1.09 m.
By courtesy of the Rijksmuseum, Amsterdam

were exclusive to court circles, the aristocracy, and the upper ranks of the bourgeoisie—there were also candlesticks. Their design was a later development of that used for altar candlesticks. The principle of a disk-shaped foot and a baluster shaft with a spike on top remained standard from the Middle Ages well into the 19th century, though the design of the individual components was affected by the styles current in any particular period. In Dinant and Flanders in the 15th century, for instance, the shaft began to be fashioned into the shape of a human figure. This style also became popular in Germany.

Whereas bronze sculpture reached its peak in Italy in the 15th century, monumental bronze figures were still rare in northern Europe at this time. Thus, the full-length equestrian statue of St. George (1373) on Hradčany Castle in Prague, which was cast by Martin and Georg von Klausenberg, did not set a trend, though rich figure decoration is often found on large fonts dating from the 13th to the 15th century. Engraved tombstones and entire tombs based on earlier traditions continued to be made until the late Gothic era (the beginning of the 16th century), as did tabernacles and lecterns.

The intellectual content of the Renaissance and the styles it engendered entered the world of the northern sculptors in the second decade of the 16th century. The Nürnberg workshop run by the Vischer family, which had been flourishing since the 15th century, continued to work in the late Gothic style until it had completed the St. Sebald's Shrine (1516), but shortly after this the style and intellectual concepts current in Italy were adopted by bronze casters in northern art centres as well. Small-scale bronze sculpture was particularly popular at this time, though some workshops were still casting monumental bronzes as late as the 18th century. (H.-U.H.)

England. Casting in bronze reached high perfection in England during the Middle Ages. The most remarkable of the sanctuary rings, or knockers, that exist at Norwich and elsewhere is that on the north door of the nave of Durham cathedral, from the first half of the 12th century.

The Gloucester candlestick

The Gloucester candlestick, in the Victoria and Albert Museum, London, displays the power and imagination of the designer as well as an extraordinary manipulative skill on the part of the founder. According to its inscription, this candlestick, which stands about two feet (60 centimetres) high and is cast in bell metal and gilded, was made for Abbot Peter (the cathedral was originally an abbey church), who ruled early in the 12th century. While the outline is carefully preserved, the ornament consists of a mass of figures of monsters, birds, and men, mixed and intertwined to the verge of confusion. As a piece of casting, it is a triumph of technique.

There remain in England 10 effigies cast in bronze over a period of two centuries (1290–1518), among them some of the finest examples of figure work and metal casting to be found in Europe. In several instances, particulars for the contracts of the tombs survive, together with the names of the artists who designed and made them. The earliest examples are the effigies of Queen Eleanor, wife of Edward I (1290), and that of Henry III (1291), both in Westminster Abbey (Figure 159). They are the work of William Torel, goldsmith of London; and it is evident that they are the first English attempt to produce large figures in metal. Torel cast his large figures by the same process (lost-wax) he had employed for small shrines and images.

Monumental brasses were exceedingly numerous in England, where some 4,000 still exist. From the 13th through the 16th centuries, in France, northern Germany, Belgium, and particularly England, it became the vogue to set into the stone slab covering a floor tomb a brass plate engraved with the figure of the deceased. The art began in Flanders and Germany, and many of the English brasses were of foreign origin; in some cases, brass sheets were imported and engraved by English artists. The manufacture of unornamented brass plates centred chiefly at Cologne. The oldest English brass in existence is that of Sir John D'Abernon (died 1277) at Stoke d'Abernon, Surrey. Traces can still be seen in many brasses of the colours that originally enlivened them. (S.V.G.)

France. In France, bronze was common from the late

J.R. Freeman & Co. Ltd.

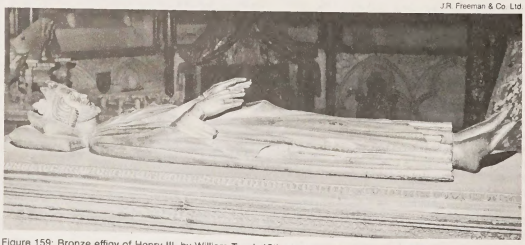


Figure 159: Bronze effigy of Henry III, by William Torel, 13th century. In Westminster Abbey, London. Length 1.50 m.

French ormolu

16th century through the 17th, 18th, and 19th centuries, and it is still popular with French sculptors today. Eighteenth-century artists made use of ormolu, or fire gilding, for bronze articles such as candlesticks, brackets, and mounts for furniture. This tradition continued in France and, to a lesser extent, in the areas under French influence, until the Empire period in the early 19th century. Subdued classical designs executed in simple brass or in bronze, generally ungilded, are typical of the period following the reign of Napoleon.

The second quarter of the 19th century and, with it, the onset of industrialization, brought about a decline in bronze casting, as it did in all spheres of craftsmanship. The age of steel production now began. At the end of the 19th century, during the Art Nouveau period, attempts were made to revive the craft of casting bronze articles; but these did not have any lasting success. Bronze continued to be used by a few individual sculptors, however, throughout the 19th century and into the present day.

(H.-U.H.)

SILVER AND GOLD

Antiquity. *Pre-Mycenaean.* Gold and silver and their natural or artificial mixture, called electrum or white gold, were worked in ancient Greece and Italy for personal ornaments, vessels, arrows and weapons, coinage, and inlaid and plated decoration of baser metals.

Aegean lands were rich in precious metals. The considerable deposits of treasure found in the earliest prehistoric strata on the site of Troy are not likely to be later than 2000 BC. The largest of them, called Priam's Treasure, is a representative collection of jewels and plate. Packed in a large silver cup were gold ornaments consisting of elaborate diadems or pectorals, six bracelets, 60 earrings or hair rings, and nearly 9,000 beads. Trojan vases have bold and simple forms, mostly without ornament; but some are lightly fluted. Many are wrought from single sheets of metal. The characteristic handle is a heavy rolled loop, soldered or rivetted to the body. Bases are sometimes round or pointed, sometimes fitted with separate collars but more often slightly cupped to make a low ring foot. One oddly shaped vessel in gold is an oval bowl or cup with a broad lip at each end and two large roll handles in the middle. The oval body has Sumerian affinities. A plain, spouted bowl in the Louvre is a typical specimen of goldsmith's work from pre-Mycenaean Greece. The scarcity of precious metals points to lack of wealth as prime cause of the artistic backwardness of these regions. Silver seems to have been more plentiful in the Greek islands; but only a few simple vessels, headbands, pins, and rings survive.

Trojan vases

Minoan and Mycenaean. A profusion of gold jewelry was found in early Minoan burials at Mŏkhlos and three silver dagger blades in a communal tomb at Kumasa. Silver seals and ornaments of the same age are not uncommon. An elegant silver cup from Gournia belongs to the next epoch (Middle Minoan I, *c.* 2000 BC). Numerous imitations of its conical and carinated (ridged) form in clay and of its metallic sheen in glazed and painted decoration prove that such vessels were common. Minoan plate and jewelry are amply represented in the wealth of mainland tombs at Mycenae and Vaphio. The vases from Mycenae are made indifferently of silver, gold, and bronze; but drinking cups, small phials, and boxes are generally made only of gold; and jugs are made of silver. Much funeral furniture is gold, notably masks that hid the faces or adorned the coffins of the dead. It has been thought that small gold disks, found in prodigious quantities (700 in one grave), were nailed on wooden coffins; but they may have been sewn on clothes. They are impressed with geometrical designs based on circular and spiral figures, stars and rosettes, and natural forms such as leaves, butterflies, and octopods. Smaller bossed disks bearing similar patterns may be button covers. Models of shrines and other amulets are also made of gold. A splendid piece of plate is a silver counterpart of a black steatite, or soapstone, libation vase from Knossos in the form of a bull's head, with gold horns, a gold rosette on the forehead, and gold-plated muzzle, ears, and eyes. (The gold here and in other Mycenaean plating is not laid on the silver but on inserted copper strips.)

Funeral furniture

Gold cups from Mycenae are of two main types: plain curved or carinated forms related to the silverware and pottery of Troy and embossed conical vessels of the Minoan tradition. Some of the plain pieces, such as the so-called Nestor's cup (Figure 160), have handles ending in animals, which bite the rim or peer into the cup. The em-

Alison Frantz



Figure 160: Mycenaean gold cup, the so-called Nestor's cup, decorated with birds on handles, from the royal graves at Mycenae, Greece, *c.* 1600–1500 BC. In the National Archaeological Museum, Athens. Height, excluding handles, 14.5 cm.

bossed ornament consists of vertical and horizontal bands of rosettes and spiral coils and of floral, foliate, marine, and animal figures. The designs are beaten through the walls and are consequently visible on the insides of most of the vessels; but the finest examples of their class, two gold cups from the Vaphio tomb near Sparta, have a plain gold lining that overlaps the embossed sides at the lip. The reliefs on the Vaphio cups represent men handling wild and domesticated cattle among trees in a rocky landscape. (Steatite vases carved with similar pictorial reliefs were evidently made to imitate embossed gold.) The handles show the typical Minoan form: two horizontal plates rivetted to the body at one end and joined at the other by a vertical cylinder.

Weapons

Cretan and mainland tombs have produced many examples of weapons adorned with gold. Modest ornaments are gold caps on the rivets that join hilt and blade, but the whole hilt is often cased in gold. An example from Mycenae has a cylindrical grip of openwork gold flowers with lapis lazuli in their petals and crystal filling between them; the guard is formed by dragons, similarly inlaid. The most splendid Mycenaean blades are bronze inlaid with gold, electrum, silver, and niello. Here again the work is done on inserted copper plates. This kind of flat inlay seems to have been originally Egyptian; it occurs on daggers from the tomb of Queen Aah-Hotep, which are contemporary with the Mycenaean (*c.* 1600 BC). Moreover, it is significant that two of the Mycenaean designs have Egyptian subjects (cats hunting ducks among papyrus clumps beside a river in which fish are swimming), though their style is purely Minoan. Another blade bears Minoan warriors fighting lions and lions chasing deer. A dagger from Thira has inlaid ax heads; one from Argos, dolphins; and fragments from the Vaphio tomb show men swimming among flying fish. These are masterpieces of Minoan craftsmanship. In the long, subsequent decadence of the Mycenaean age, however, there seems to have been no invention, and later pieces of goldsmiths' work repeat conventional forms and ornaments. (E.J.F./M.C.R.)

Iran. The Persians have been skillful metalworkers since the Achaemenid period (559–330 BC), when they were already acquainted with various techniques such as chasing, embossing, casting, and setting with precious stones. Statuettes of gold and silver are known from the 5th century BC, and vessels of silver and gold from this time

take the form of phials, conical cups, vases, and rhyta (drinking cups in the shape of an animal's head). The Oxus treasure in the British Museum and the Susa find in the Louvre, Paris, are good examples of such work. During the Parthian period (247 BC–AD 224), silverwork and goldwork was strongly influenced by Hellenistic predilection for richly decorated bowls and dishes. The zenith of old Iranian metalwork, however, was reached during the Sāsānid period (AD 224–651), when craftsmen achieved great variety in shape, decoration, and technique. Drinking vessels (stem cups and cups with handles), ewers, oval dishes, platters, and bowls are the dominant forms; hunting scenes, drinking scenes, and animals are represented

Sāsānid craftsmanship



Figure 161: Persian embossed silver bowl showing a king slaying lions, Sāsānid period c. AD 224–651. In the British Museum. Diameter 24.1 cm.
By courtesy of the trustees of the British Museum

in high relief (Figure 161). The patterns were cut out of solid silver or made separately in sheets and then soldered to the vessel. From this time onward cloisonné enamel was used for jewelry. (B.V.Gy.)

Greek and Etruscan. The period of transition from the Bronze to the Iron Age, when Aegean external relations were violently interrupted, was not favourable either to wealth or art; and the only considerable pieces of plate that have come from Greece are embossed and engraved silver bowls made by Phoenicians. Most of them bear elaborate pictorial designs of Egyptian or Assyrian character and are evidently foreign to Greece; but some simpler types, decorated with rows of animals in relief or wrought in the shape of conventional flower bowls, can hardly be distinguished from the first Hellenic products. A severe and elegant silver bowl in the Metropolitan Museum of Art represents the flower type in its finest style. It is cast and chased and probably belongs to the 5th century BC.

Silver vases and toilet articles have been found beside the more common bronze in Etruscan tombs; for example, a chased powder box of the 4th century BC in the Metropolitan Museum of Art. Bronze reliefs of an archaic chariot in the same collection have their opulent counterparts in some hammered silver and electrum fragments in London, Munich, and Perugia. The electrum details are attached with rivets.

Roman silver relief

Roman. About the 4th century BC, the fashion of ornamenting silver vessels with relief was revived; and this type of work, elaborated in the Hellenistic Age and particularly at Antioch and Alexandria, remained the usual mode of decoration for silver articles until the end of the Roman Empire.

The scholar Pliny the Elder (1st century AD) names Greek silversmiths whose work was valued highly at Rome and laments the disappearance of the art in his own day. He must refer only to its quality, for Roman silverware has been abundantly preserved. Many rich hoards in modern collections were buried by design during the calamitous last centuries of the ancient world; and the most sumptuous, the Boscoreale treasure (mostly in the Louvre), was accidentally saved by the same volcanic catastrophe that destroyed Herculaneum and killed Pliny in AD 79 (Figure 162). A slightly smaller hoard found at Hildesheim (now in Berlin) also belongs to the early empire. The acquisition and appreciation of silver plate was a sort of cult in Rome. Technical names for various kinds of reliefs were in common use (*emblemata, sigilla, crustae*); weights were recorded and compared and ostentatiously exaggerated. Large quantities of bullion came to Rome with the spoils of Greece and Asia in the 2nd century BC; and Pliny says that even in republican times there were more than 150 silver dishes of a hundredweight apiece in the city. (Weights of vessels are often marked on their bases.)

Cups and jugs

Cups and jugs of Augustan style are usually covered with ornament in high relief. The subjects are very diverse: historical, mythological, and mystic scenes, formal and naturalistic designs of flowers and foliage, graceful studies of animals and birds. Some cups and jugs have conventional fluting, petals, or gadroons (ornamental bands embellished with continuous patterns); Bacchic masks; and embossed or engraved wreaths, gilt or inlaid with niello. Silver and niello inlay was commonly applied to bronze plates. A singular type of silver bowl (*patera clipeata*) has a central ornament in high relief or even in the round; the ornament frequently contains a portrait bust. In time the ornament was restricted; and later Roman plate is plain with narrow border friezes, small central medallions, and handles embossed in low relief. One of the very few gold pieces that survive, a shallow bowl found at Rennes (Bibliothèque Nationale), is exceedingly elaborate. It measures 10 inches across and weighs 46 ounces. The central medallion and its surrounding frieze contain scenes of a drinking contest between Bacchus and Hercules; between the frieze and the edge of the bowl is a row of 16 gold coins, each framed in a foliate wreath. The coins range from Hadrian to Caracalla. In the same collection are several examples of very large silver plates (*clipei* or *missoria*), in which the whole field is embossed with mythological or historical subjects. The largest (called the Shield of Scipio) is 28 inches in diameter and weighs 363 ounces.

(E.J.F./M.C.R.)

Early Christian and Byzantine. The earliest Christian silverwork closely resembles the pagan work of the period in its naturalistic grace, ornament, and use of the traditional techniques of embossing and chasing. Even the subject matter is sometimes classical: the late 4th-century

By courtesy of the Musee du Louvre, Paris.
Cliche Musees Nationaux



Figure 162: Roman silver pitcher, from the villa at Boscoreale, near Pompeii, Italy, c. 1st century BC. In the Louvre, Paris. Height 25 cm.



Figure 163: Early Christian marriage casket of Projecta and Secondus, embossed silver, partially gilded, from the Esquiline treasure, Rome, c. 400. In the British Museum. Length 60.33 cm.

By courtesy of the trustees of the British Museum

marriage casket of Projecta and Secondus (Figure 163), part of the Esquiline treasure found at Rome (British Museum), is decorated with pagan scenes; and only the inscription shows that it was made for a Christian marriage. Among the few pieces with Christian subjects are small Roman cruets (condiment bottles) from Taprain, Scotland (Royal Scottish Museum, Edinburgh, and the British Museum), and a small pyx (casket for the reserved Host) from Pola, Yugoslavia (Kunsthistorisches Museum, Vienna).

Most of the silver of the latter part of the period has been found in the Christian East—in Syria, Egypt, Cyprus, Asia Minor, and Russia—and is mostly "church" plate (chalices, censers, candlesticks, and bowls and dishes probably used to hold the eucharistic bread). Secular plate was also decorated with religious subjects—for example, dishes depicting the life of David (Cyprus Treasure, Cyprus Museum, Nicosia, and Metropolitan Museum); both dishes and vessels were produced with pagan subjects—for example, the Concesti amphora and the Silenus Dish (both in the Hermitage, Leningrad). The figure style is often harder and flatter than previously, characterized by strictly frontal positions and symmetry. The techniques of chasing and embossing still predominated, but abstract patterns and Christian symbols inlaid in niello were used increasingly. The appearance of imperial "control stamps," early forerunners of hallmarks, show most of this material to be of the 6th and 7th centuries. It is not known which cities were important centres of production; but the Eastern capital, Constantinople, must have been foremost among them.

Religious decoration

Of work in gold of the earliest Christian period, only personal jewelry has survived; but from the 6th and 7th centuries onward other pieces are also extant. Among the most important of the latter are votive crowns and crosses offered to churches in Spain and Italy by royal patrons. The finest of these pieces are those found in Guarrazar in Toledo Province (National Archaeological Museum, Madrid, and Musée de Cluny, Paris), inlaid with garnets and jewels; the cross of King Agilulf (cathedral of Monza, Italy); and a pair of gold book covers inscribed by Queen Theodolinda (cathedral of Monza, Italy). The book covers are set with pearls, gems, and cameos and decorated with gold cloisonné work inlaid with garnets, a popular style among the Germanic peoples. Inlaid cloisonné jewelry reached an especially high standard of workmanship in Britain, as is shown by a purse lid, a sword, and jewelry from the cenotaph (monument honouring a dead person whose body lies elsewhere) to a 7th-century East Anglian king discovered at Sutton Hoo, Suffolk (British Museum). Major works in silver and gold were also produced in the northern Hiberno-Saxon school and in the service of the Celtic Church; work in precious metal, such as the buckle on the Moylough belt reliquary and the Ardagh Chalice in the National Museum of Ireland, Dublin, displays a masterly synthesis of the northern arts and humanist Mediterranean tradition.

Middle Ages. *Carolingian and Ottonian.* The earliest works of the Carolingian renaissance, made in the last quarter of the 8th century, resemble Hiberno-Saxon art of the 8th century in their abstract treatment of the human figure, their animal ornament, and their use of niello and "chip-carving" technique; examples are the Tassilo Chalice (Kremsmünster Abbey, Austria) and the Lindau Gospels book cover (Pierpont Morgan Library, New York City). From about 800 onward, however, the influence of the Mediterranean tradition gained strength at Charlemagne's court at Aachen and later spread through the whole empire. Triumphal arches (now lost) given by the Emperor's biographer Einhard to Maastricht cathedral were typical of this movement; miniature versions nine inches (22 centimetres) high of great marble triumphal arches of antiquity, they were embossed in silver with Christian subjects. The bulk of work in precious metals that survives from the Middle Ages is ecclesiastical: golden altars, like that of S. Ambrogio in Milan (*c.* 850), where scenes from the life of Christ and St. Ambrose are framed by panels of cloisonné enamel and filigree (openwork); and reliquaries and book covers in gold and silver, set with gems and decorated by embossed figures and scenes, such as the cover of the Codex Aureus of St. Emmeram (*c.* 870; Bayerische Staatsbibliothek, Munich). These pieces testify to the magnificence of Carolingian work, the techniques of which were to dominate the goldsmith's craft until the 11th century.

Ecclesiastical work

Patronage throughout this period was mainly in the hands of the emperors and great princes of the church; and the form of liturgical plate and reliquaries, altar crosses, and the like underwent no fundamental change; Ottonian work of the later 10th and 11th centuries can be distinguished from that of the 9th only in the development of style. For example, the larger, more massive figures, with their strict pattern of folds, on the golden altar (*c.* 1023) given by Henry II to Basel Minster (Musée de Cluny, Paris), are markedly different from the nervous, elongated figures of the Carolingian period.

Romanesque. In the 12th century the church supplanted secular rulers as the chief patron of the arts, and the work was carried out in the larger monasteries. Under the direction of such great churchmen as Henry, bishop of Winchester, and Abbot Suger of Saint-Denis, near Paris, a new emphasis was given to subject matter and symbolism.

Craftsmen were no longer anonymous; work by Roger of Helmarshausen, Reiner of Huy, Godefroid de Claire (de Huy), Nicholas of Verdun, and others can be identified; and the parts they played as leaders of the great centres of metalwork on the Rhine and the Meuse are recognizable. Their greatest achievement was the development of the brilliant champlevé enamelling, a method that replaced the earlier cloisonné technique. Gold and silver continued to be used as rich settings for enamels; as the framework of portable altars, or small devotional diptychs or triptychs; for embossed figure work in reliquary shrines; and for liturgical plate.

The masterpieces of the period are great house-shaped shrines made to contain the relics of saints; for example, the shrine of St. Heribert at Deutz (*c.* 1160) and Nicholas of Verdun's Shrine of the Three Kings at Cologne (*c.* 1200). In the latter, the figures are almost freestanding, and in their fine, rhythmic draperies and naturalistic movement they approach the new Gothic style.

Shrines and reliquaries

Gothic. The growing naturalism of the 13th century is notable in the work of Nicholas' follower Hugo d'Oignies, whose reliquary for the rib of St. Peter at Namur (1228) foreshadows the partly crystal reliquaries in which the freestanding relic is exposed to the view of the faithful; it is decorated with Hugo's particularly fine filigree and enriched by naturalistic cutout leaves and little cast animals and birds.

The increasing wealth of the royal courts, of the aristocracy, and, later, of the merchants led to the establishment of secular workshops in the great cities and the foundation of confraternities, or guilds, of goldsmiths and silversmiths, the first being that of Paris in 1202.

As in architecture, monumental sculpture, and ivory carving, the lead held by Germany and the Low Countries during the Romanesque period now passed to France. Architectural forms continued to be the basis of design in precious metal; the silver shrine of St. Taurin at Évreux (*c.* 1250), for example, is a Gothic chapel in miniature,

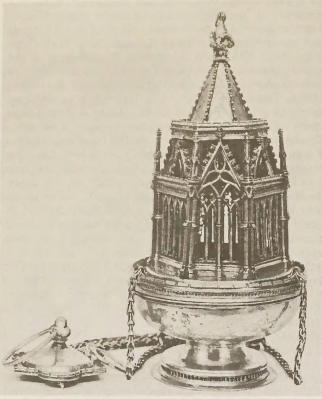


Figure 164: Ramsey Abbey censer, cast, embossed, and gilt silver. English Gothic, 14th century. In the Victoria and Albert Museum, London. Height 27.6 cm.

Crown Copyright Victoria and Albert Museum, London

with saints under pointed arches, clustered columns, and small turrets. In England, the few pieces that survived the dissolution of the monasteries in the 16th century follow the same architectural pattern. Notable examples are the 14th-century Ramsey Abbey censer (Figure 164) and the magnificent crosier made for William of Wykeham (New College, Oxford). Germany first produced work in the Gothic style in the second half of the 14th century with a large Gothic head reliquary of Charlemagne and the splendid "Three-Tower" reliquary, both still at Aachen. In Italy, despite the undercurrent of classical taste, the Gothic style predominated in the 14th century, especially at Siena; it was also probably in Italy around 1280 that basse-taille enamel—a technique in which intaglio relief carving in the metal below its surface is filled with translucent enamel—originated, whence it spread rapidly through the upper Rhine region to France and England. The Parisian school of enamellers predominated in the latter half of the 14th century. For the first time, enough secular plate survives to show that it equalled the ecclesiastical in opulence: two fine pieces are the Royal Gold Cup made in Paris around 1380 (British Museum) and the so-called King John's Cup, probably English work of around 1340 (King's Lynn, Norfolk).

Basse-taille enamel

The late Gothic period produced court treasures such as the "Goldenes Rössel" (1403; Stiftskirche, Altötting, West Germany), and the Thorn reliquary (British Museum), both early 15th century. There was also an increased output of secular silver because of the rise of the middle classes; the English mazers (wooden drinking bowls with silver mounts) and the silver spoons with a large variety of finials are examples of this more modest plate. Numerous large reliquaries and altar plate of all kinds were still produced. At the end of the Middle Ages the style of these pieces and of secular plate developed more distinctive national characteristics, strongly influenced by architectural style: in England, by the geometric patterns of the Perpendicular; in Germany, by heavy and bizarre themes of almost Baroque exuberance; and in France, by the fragile elegance of the Flamboyant.

"Hallmarking"

The purity standards of silver became rigorously controlled, and "hallmarking" was enforced; the marking of silver in England, especially, was carefully observed.

(P.E.L.)

Islām. The use of gold and silver in Islāmic lands was limited because it was forbidden by the Qur'ān, and although the prohibition was often ignored, the great value of such objects led to their early destruction and melting down. Islāmic jewelry of the early period is therefore of extreme rarity, represented only by a few items, such as buckles and bracelets of the Fātimid and Mongol periods and such pieces as the Gerona silver chest (akin to similar ivory coffers) in Spain and the Berlin silver tankard of the 13th century, with embossed reliefs of Sāsānian animal friezes.

(H.Go.)

Renaissance to modern. *16th century.* Italian goldsmiths preceded the rest of Europe in reverting to the style of Roman antiquity; but in the absence of antique goldsmiths' work, vases of marble or bronze had to serve as models. Goldsmiths often worked from very free interpretations of the antique made by artists in other media. Many of these designs but very few of the actual pieces have survived; the most famous is an enamelled gold saltcellar (Kunsthistorisches Museum, Vienna) made for Francis I by the celebrated Florentine Benvenuto Cellini. In the second half of the 16th century many gifted Italian and immigrant goldsmiths worked at the court of Cosimo I, grand duke of Tuscany, specializing in vessels of hardstone mounted in enamelled and jewelled gold; their work is well represented in the Museo degli Argenti in the Pitti Palace, Florence, and in the Kunsthistorisches Museum; similar work was done by the Sarachi family in Milan.

Little French goldwork is extant, and most of the surviving material is in the Galerie d'Apollon in the Louvre. Among the most sumptuous pieces are a sardonyx (a type of onyx) and gold ewer, the gold St. Michael's Cup (both at the Kunsthistorisches Museum), and a sardonyx-covered

Archivo Mas, Barcelona

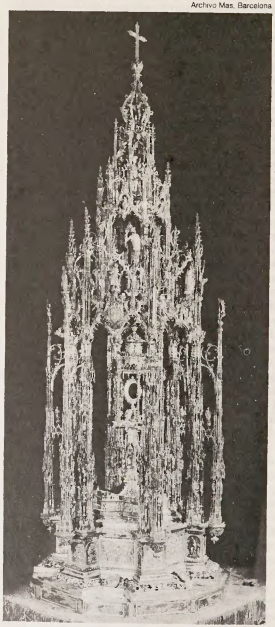


Figure 165: *Custodia* of goldwork, silverwork, and enamel work (1515–23), by Enrique de Arfe. In the Toledo Cathedral, Spain. Height 2.50 m.

cup in the Louvre, all of which display northern features. The massive plate of the Ordre du Saint-Esprit (Louvre), dating from 1581–82, is of quite individual character; and an enamelled gold helmet and shield of Charles IX (1560–74) in the Louvre have no parallel either for quality or opulence.

In other parts of Europe, goldsmiths clung to Gothic forms until well into the first half of the century, especially in the provincial towns. Immensely rich in ecclesiastical silver, Spain has little early domestic silver; Spanish silversmiths, *platería*, gave their name to the heavily ornamented style of the period, Plateresque. Using precious metal from the New World, goldsmiths such as Enrique and Juan de Arfe produced vast containers for the Host known as *custodia* (Figure 165). The most important Portuguese work, the Belém monstrance, created by Gil Vicente in 1506 for Belém Monastery near Lisbon, is still Gothic in style; later, Portugal developed its own style, related to Spanish work but not copied from it.

Spanish and Portuguese goldsmiths

Some of the finest 16th-century goldsmiths' work was executed in Antwerp and elsewhere by such Flemish goldsmiths as Hans of Antwerp, goldsmith to Henry VIII, and Jacopo Delfe, called Biliverti, goldsmith to Cosimo I. The Flemish masters showed particular sympathy for the Mannerist style, derived from Italy but transformed by such native engravers as Cornelis Bos and Cornelis Floris. By about 1580, Dutch goldsmiths had begun to rival the Flemish; the van Vianen family of Utrecht won international renown, especially Adam, who excelled at embossing, and his brother Paulus, who worked in Italy, Munich, and in the workshop of Rudolph II at Prague.

The principal centres in the north were Nürnberg and Augsburg, the former particularly notable for the exuberant Mannerism of the Jamnitzer family, the latter for its ebony caskets with silver-gilt mounts. Many German princes, especially the dukes of Bavaria, maintained their own court workshops. Production was on a vast scale, and great quantities survive. Characteristic German forms are columbine cups (the trial piece for entry into the Nürnberg Goldsmith's Guild) and standing cups such as the Diana Cup (Figure 166) by Hans Petzolt.

By courtesy of the Kunstgewerbemuseum, West Berlin



Figure 166. Diana Cup, silver standing cup by Hans Petzolt, Nürnberg, *c.* 1610. In the Kunstgewerbemuseum, West Berlin. Height 80.01 cm.

England is rich in 16th-century secular silver, but church plate was mostly destroyed during the Reformation. The Renaissance style, introduced by the painter Hans Holbein the Younger, who designed vessels for the court, follows that of the Low Countries and Germany. Certain individual forms also were produced, such as standing saltcellars with tiered covers and "steeple" cups, which had a tall finial on the cover.

Baroque. In the first half of the 17th century Dutch goldsmiths, such as the van Vianens and, later, Johannes Lutma the Elder of Amsterdam, developed a fleshy form of ornament known as auricular, which became common in northern Europe, including England—where Christian van Vianen worked as court goldsmith to Charles I—and Germany—where the Thirty Years' War (1618–48) reduced both the quantity and quality of production. After midcentury, bold Dutch floral ornament—usually embossed in thin metal, as though the pieces were for display rather than use—was characteristic and influential. France, however, undoubtedly led fashion with its state workshops at the Gobelins, the refined French acanthus ornament contrasting sharply with the coarser Dutch designs. Since Louis XIV melted the royal plate to pay his troops, no French work of this period remains; but its quality is demonstrated in the work of the Huguenot silversmiths who left France after the revocation of the Edict of Nantes in 1685. Mostly provincials, they brought new standards of taste and craftsmanship wherever they settled—particularly in England, where the foremost names of the late 17th and earlier 18th centuries were of French origin: Pierre Harache, Pierre Platel, David Willaume, Simon Pantin, Paul de Lamerie, Paul Crespin, to mention but a few.

Silver furniture

Silver furniture, a feature of the state rooms at Versailles, became fashionable among kings and noblemen. It was constructed of silver plates attached to a wooden frame; and each suite contained a dressing table, a looking glass, and a pair of candlestands. In France such furniture did not survive the Revolution; but much remains in England, Denmark, Germany, and Russia.

After the Thirty Years' War, Germany did not regain its eminence; even the enamelled goldwork from the court workshops at Prague and Munich, which became larger and more ostentatious in colour, was inferior in design and finish. In Scandinavia, particularly Sweden, goldsmiths evolved forms of beakers and tankards showing strong German influence. Spanish silver was of massive architectural design, oval champlevé enamelled bosses being set at intervals over the surface of the larger pieces. The few extant Italian pieces suggest that the goldsmiths worked their material with the skill of sculptors.

18th century. Early 18th-century English work combined functional simplicity with grace of form, while the work of Dutch and German goldsmiths is in a similar style but of less pleasing proportions. The preeminence of the English work, however, is due to the destruction of all but a fraction of French silver of the same period; for what survives is outstanding in originality of design and fineness of finish. The superiority of French work lay in its excellence of design and the high quality of the cast and chased work. Where other goldsmiths worked in embossed metal, the French modelled and cast their ornament and then applied it—a technique that consumed much more of the precious material.

Superiority of French work

In France, provincial goldsmiths competed successfully with those of the capital; but in England all the best artists went to London. In the early 1730s the French Rococo style was imported to England and adopted by goldsmiths of both Huguenot and English descent, one of the latter being Thomas Heming, goldsmith to George III. English silver in the 18th-century classical style of Robert and James Adam is of unequal merit owing to the use of industrial methods by some large producers.

In France, Robert Auguste created pieces of great refinement in the Neoclassical style, which was copied in Turin and in Rome, for example, by L. Valadier. A notable

workshop was founded in Madrid in 1778 by D. Antonio Martínez, who favoured severely classical designs. In both the northern and southern Netherlands, local production followed French precept, but more individuality survived in Germany. In Augsburg, excellent table silver was produced, but more important were the pictorial panels embossed in the highest relief by members of the Thelot family and the silver furniture made by the Billers and the Drentwetts. At Dresden, Augustus II the Strong established under Johann Melchior Dinglinger a court workshop that produced jewels and enamelled goldwork unequalled since the Renaissance; and the gold snuffboxes made by Johann Christian Neuber rivalled those of the Parisian goldsmiths. (J.F.Ha.)

Colonial America. Silversmithing in the New World in the colonial period is more or less derivative from Europe and England. In North America it was first brought to New England by English craftsmen in the 17th century. The most important centres were Boston, Newport, New York City, Philadelphia, Baltimore, and Annapolis. Outstanding collections include the Mabel Brady Garvan collection at Yale University and those in the Boston Museum of Fine Arts, the American Wing of the Metropolitan Museum of Art, and in the Philadelphia Museum of Art. North American colonial silver is distinguished for its simplicity and graceful forms, copied or adapted from English silver of the period. On the other hand, the colonial silver of Mexico, Brazil, Colombia, Peru, Chile, and Bolivia, while European in concept, shows a blending of Iberian designs and forms, with indigenous influences that trace back to pre-Hispanic times. Most of these relics survive in churches as sacramental vessels; but there are some notable private collections. (D.T.E.)

Colonial silver of Mexico and Central and South America

19th century. The Napoleonic adventure brought French fashions back into prominence, and the Empire style was widely followed on the Continent. In England the Regency goldsmiths, of whom Paul Storr was the foremost, created their own more robust version of the Empire style. Perhaps the most impressive monument of the period is a service made in Lisbon between 1813 and 1816 and presented to the Duke of Wellington for his liberation of Portugal (now in Apsley House, London).

By midcentury most of the earlier styles had been revived fleetingly and a recognizable Victorian style evolved, based on details drawn from diverse sources. Craftsmanship was at its best, but the design of domestic silver was derivative and selective, while that of presentation pieces strove too consciously for naturalistic effect. In the latter half-century the craft became an industry and the goldsmith a factory worker. In this respect Matthew Boulton was the great pioneer: his Soho manufactory near Birmingham, which dominated the British "toy" industry from the 1770s, produced high-quality steel buckles, buttons, coins, sterling silver, and Sheffield plate, establishing standards of design and of factory management and welfare services that rivalled those of the 20th century. At the end of the 19th century, standards deteriorated, and a second pioneering movement started—the craft revival associated with William Morris and the Art Nouveau style (see below *Modern*), which led to the production of original pieces, some of highly mannered design. In England the most interesting work was done by the sculptor Sir Alfred Gilbert, who, following the lead of William Burges, the architect and designer, combined silver with ivory and semiprecious stones in romantic confections. (J.F.Ha.)

The craft as industry

Modern. The structure of trade, following the drastic social changes that have taken place since 1914, is similar in all industrial countries. A few artist-craftsmen maintain independent studio workshops, producing commercially unprofitable but artistically significant work. Many of them also teach in art schools or work part-time in factories as industrial designers. Factories using modern equipment—for example, stamping, pressing, spinning, casting, and mechanical polishing—account for nearly all the financial turnover but seldom break new ground artistically. Retail shops buy stock almost entirely from the factories and wholesalers and usually sell it anonymously. Thus, the evolution of style is impeded by the cost of new machinery; by the natural caution of wholesalers and retailers; by the buying public, which prefers precious ornaments to be timeless; and by the consideration that buying is an investment for value rather than for beauty. In consequence, the most lively designs are often those for costume jewelry; and the best modern work usually has been on a tiny scale, making little impact on the trade.

Decline of creativity

In Paris, designs by René Lalique inspired Art Nouveau, which spread to Belgium and then through Europe and the United States. In Moscow, Peter Carl Fabergé set a superb standard of craftsmanship for small ornaments. In Denmark, Georg Jensen, with Johan Rohde and others, achieved not only an individual Danish style but built up several factories with retail outlets across the world, thus proving that good modern design in silver and jewelry need not be confined to artists' studios (Figure 167); their

By courtesy of Jensen & Co., Copenhagen

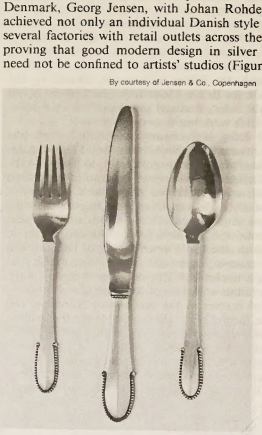


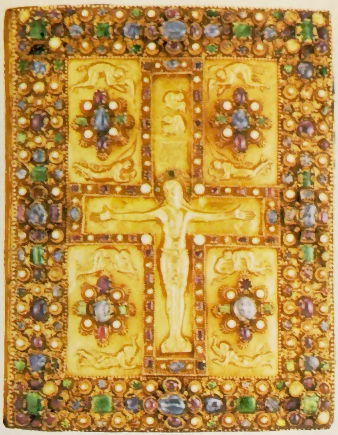
Figure 167: Sterling silver knife, fork, and spoon, designed by Georg Jensen, Copenhagen, 1916.

influence spread throughout Scandinavia. In the 1960s only Germany approached Scandinavia in the number and quality of its artist-craftsmen; WMF (Württembergische Metallwarenfabrik) at Geislingen is probably the biggest silverware factory in Europe. In England, notable for the most varied work, the Worshipful Company of Goldsmiths has helped a vigorous group of designers to emerge since 1945, including Gerald Benney, Eric Clements, David Mellor, John Donald, and Andrew Grima. (G.McK.H.)

PEWTER

In its pure form, tin is far from suitable for making into implements because it is too brittle for casting successfully and is not easy to melt down. For this reason it has always been alloyed with certain other metals, mainly lead, in the proportion of 10:1, or copper, alloyed about 100:4, to make what is known as pewter. In medieval Germany, the municipal authorities and the guilds laid down permissible ratios to be used for tin alloys. The authorities also kept an eye on the pewterers and their products to make sure that regulations were adhered to. So that pewter ware could be kept under constant surveillance, a system was worked out whereby every single article had to be marked by one, two, or more hallmarks, or "touches." The first decrees of this kind to be issued in Germany date from the 14th century. In France and England, written sources refer to the pewterer's obligation to hallmark his wares from the end of the 15th century onward. These regulations do not seem to have been followed very closely in practice, for pieces surviving from the period before 1550 rarely have the regulation marks. In the second half of the 16th century, however, which was the golden age of pewter, almost all work began to be clearly marked. This means that modern collectors have a good chance of being able to identify their pieces.

Pewter ware is cast in molds. It is not suitable for chas-



Book cover of the Lindau Gospels (MS. 644, fol. 115v), chased gold with pearls and precious stones, Carolingian. In the Pierpont Morgan Library, New York City. 27 × 35 cm.

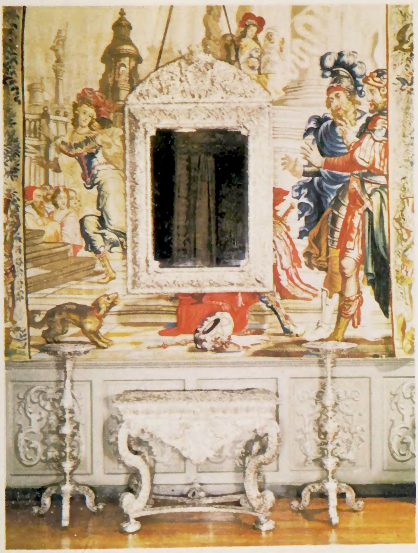


English silver tureen with the Cavendish arms by Paul Storr, 1820–21. In Chatsworth House, Derbyshire. Height 49 cm.

Silver and gold work



Persian vase in the form of a fish, gold sheet decorated with incised lines, details of eyes and mouth in repoussée, Achaemenid period, 5th–4th century BC. In the British Museum. Length 24.2 cm.



Furniture in the king's bedroom, Knole House, Kent, England, silver on wood, 17th century. Height of table 61 cm.



Mask of Xipe Totec, gold, cast by the "lost-wax" method, Mixtec culture, *c.* 900–1494. In the Museo Regionale, Oaxaca, Mexico. Height 7 cm.

Plate 7 By courtesy of (top left) the Pierpont Morgan Library, New York, (centre) the trustees of the British Museum, (bottom left) the National Trust, by permission of (top right) the trustees of the Chatsworth Settlement, photograph (bottom right) Ferdinand Anton

Copper, bronze, and brass



Celtic bronze shield, originally gilt, 2nd century AD. In the British Museum. Height 78 cm.



Nepalese statue of Krsna, gilded bronze with turquoise and gems, 18th century. In the Prince of Wales Museum of Western India, Bombay.



Egyptian bronze plaque damascened with gold wire, 22nd dynasty (945–*c.* 730 BC). In the Metropolitan Museum of Art, New York City. Length 19 cm.

The Chimera of Arezzo, bronze, Etruscan, 5th century BC. In the Museo Archeologico, Florence. Height 80 cm.



Plate 8 By courtesy of (top left) the trustees of the British Museum, (centre) the Metropolitan Museum of Art, New York, Rogers Fund, 1908; photographs, (top right, bottom) SCALA—Art Resource
Plate 9 By courtesy of (top left) the Smithsonian Institution, Freer Gallery of Art, Washington, D.C., (centre) the trustees of the British Museum, (bottom right) the Victoria and Albert Museum, London; photographs, (top right, bottom left) Holle Bildarchiv, Baden-Baden, (centre left) [illegible] Bildarchiv, Munich



Chinese bronze *chung*, late Chou dynasty (*c.* 1122–221 BC). In the Freer Gallery of Art, Washington, D.C. Height 67 cm.



Cast bronze baptismal font by Renier de Huy, 1107–18. In the church of Saint-Barthélemy, Liège, Belgium. Height 64 cm.



Mycenaean dagger, bronze with gold, silver, and niello, 16th century BC. In the National Archaeological Museum, Athens. Length 16.3 cm.



Syrian *pome*, or hand warmer, openwork copper with silver inlay, *c.* 13th century. In the British Museum. Diameter 18.5 cm.



Gloucester candlestick, carved and chased gilt bronze, 12th century. In the Victoria and Albert Museum, London. Height 58 cm.



Portable altar, cut-out, gilded, engraved, and incised laminated copper, attributed to Roger of Helmarshausen, *c.* 1100. In the collection of the Franciscan monastery of Paderborn, Germany. Length 31.5 cm.



Pewter jug by Paul Weise, Zittau, Germany, late 16th century. In the Victoria and Albert Museum, London. Height 52.1 cm.



Rivergod symbolizing the Enns, from the lead fountain by Georg Raphael Donner (1693–1741). Formerly in the Neuen Markt, Vienna, presently in the Österreichische Galerie, Vienna. Length 2.4 m.



Detail of a wrought-iron *reja* (choir screen) by Pedro Juan, 1668, gilded in 1764. Originally in the cathedral of Valladolid, Spain, presently in the Metropolitan Museum of Art, New York City. Height 15.8 m.



Japanese *tsuba* (sword guard), iron with openwork design, Muromachi period, 1338–1573. In the National Museum, Tokyo. Diameter 10 cm.

Pewter, iron, and lead

Plate 10: By courtesy of (top left) the Victoria and Albert Museum, London, (top right) the Österreichische Galerie, Vienna, (bottom left) the Metropolitan Museum of Art, New York, gift of the Hearst Foundation, 1956, (bottom right) Eisaku Kishida-Nippon Bijyutsu Token Kyokai

Pewter molds

ing or stamping. Molds for simple utensils such as plates, bowls, and jugs were made of clay mixed with calves' hair or of plaster, stone, or slate. From the 16th century, when pewter ware began to be decorated with relief work, molds made of brass or copper were used instead. Relief decoration can be applied by two different methods. The pewterer could either chisel the relief decoration (consisting of little scenes, figures, or decorative motifs) into the copper mold in intaglio, which enabled him to make the details as three-dimensional as he wished; or he could etch it in, which involved covering the plain copper mold with wax, scratching the decoration into it, and then allowing caustic acid to act on it. This second method resulted in a rather flat, two-dimensional relief, which is reminiscent of woodcuts in its sharp outlines and overall style; thus, the technique is known as the "woodcut style." It was common practice in Nürnberg in the last quarter of the 16th century. Pewter utensils (exclusively plates and dishes at this time) were cast in molds prepared in this manner. It was very seldom that decorative motifs were etched straight onto the pewter surface.

Another type of decoration is engraving, which involves cutting decorative motifs, figures, or inscriptions with a burin into the surface of pewter objects. The most expensive and aesthetically important pieces of engraved pewter were produced in the late Gothic period, about 1500. In the 16th and 17th centuries, engraving was common for guild articles; and in the 18th century engraved mottoes, names, dates, and motifs taken from popular art were widely used. The type of strokes used fall into three categories: long, engraved lines; dots set close together to form a pattern; and a technique known in German as *Flecheln*, in which the straight line made by the burin is broken up into a series of long or short zigzag strokes. The last method makes the design look fuller and broader and also makes it stand out more sharply. This type of decoration first appeared in the 16th century and was very popular in the 17th and 18th centuries.

After they had been cast and then turned on a lathe, many pewter articles, especially plates and dishes, were hammered. The idea was to smooth over the surface of the object and strengthen the material by means of a series of light and regular blows. Sometimes pewterers punched their wares with decorative motifs stamped close together to form a sort of frieze. This technique is known as tooling and is commonly found on bronze and silver articles. Occasionally, pewter pieces were embellished by the addition of brass fittings, such as handles, knobs, spouts, or scroll panels. But pewter ware has rarely been gilded, partly because it is difficult to make a layer of gilding adhere to the surface, partly because there seems little point in covering a material that is attractive in itself with a metal that is ostensibly more precious. This is also why pewter ware has rarely been painted.

Pewter inlay

A type of pewter inlay is found on what are known as Lichtenhain tankards. Most of these tankards were made in Lower Franconia and in Thüringia in the 18th and 19th centuries. They have wooden staves running down them, and their sides are inlaid with decorative motifs and figures made of thin sheets of engraved pewter. In the early 18th century, furniture was also occasionally inlaid with pewter. Such furniture was clearly inspired by the inlay work of the French cabinetmaker André-Charles Boulle.

Antiquity. On the whole, excavations have unearthed little pewter ware dating from antiquity, not only because it has tended to perish over the years but presumably also because it was not nearly as common as glass, bronze, silver, or clay. Excavations on the Esqueline Hill and finds from the Tiber River have produced some small pewter statuettes of divinities that may well be votive offerings. Miniature versions of household articles such as amphorae, oil lamps, and pieces of furniture were found in graves.

A number of pewter ampullae (flasks with a globular body and two handles) with inscriptions or highly stylized images or symbols date from the Early Christian period. They were sold to pilgrims and were used to hold water from the Jordan River, consecrated water, or oil. (Similar pouch-shaped ampullae reappeared in France in the 14th and 15th centuries; but unlike the early Christian examples, they are ornamented with abstract motifs rather than figure decoration.)

Middle Ages. Besides the ampullae, hundreds and thousands of pilgrim badges were sold to devout visitors to places of pilgrimage in the Middle Ages. These little plaques and *agraffes* (hat badges) were generally miniature versions of religious images worshipped at the place where they were on sale. A number of these Italian, English, French, and German pilgrim badges, dating from the 13th to the 16th century, have survived.

Instead of jewelry made of gold, silver, or precious stones, the less wealthy people of the Middle Ages wore pewter badges sewn onto their clothes or hats. The badges often took the form of amulets.

Because pewter was highly prized in all periods, damaged or old-fashioned utensils were melted down over and over again to make new ones. Thus, the earliest surviving functional objects and vessels made of pewter date from the Gothic era, though a few written sources refer to pewter being used earlier than this. Most of these documents are concerned with the question of whether communion chalices should be made of anything other than gold or silver. Pewter Communion chalices were permitted in certain periods and prohibited in others, and the church never managed to draw up an absolute ruling that applied to all religious communities.

Guild flagons

Some of the finest and most important pewter pieces ever cast were made in Silesia in about 1500. Large guild flagons of a characteristic polygonal design, only 11 of them have been preserved. Their facetted surfaces are engraved with figures of saints surrounded by interlaced foliage scrolls, arches, arcades, and other late Gothic decorative motifs. Hidden among these motifs, one sometimes finds secular scenes, some of which are downright lewd. Pewterers in the neighbouring districts of Moravia and Bohemia also made guild flagons; but theirs were cylindrical, with raised horizontal bands. The areas between the bands were generally decorated with friezelike inscriptions made up of Gothic or Gothic-style characters.

The 15th century saw the emergence of a jug set on a slender stem, easily recognizable by its disk-shaped base, surmounted by another slender stem; the main body of the vessel is generally spherical and has a long, thin neck. The municipal authorities often possessed a set of six or 12 flagons of this kind. They came back into fashion in the 17th century and were very widely used, as they had been at the beginning of the 15th century. Unfortunately, only a very few have survived from the earlier periods.

Another early type of vessel belongs to a group known as Hanseatic tankards. These tankards have a heavy-looking, potbellied body set on a shallow circular base and a slightly convex lid. They were used in the coastal regions of Germany—that is, along the North Sea and Baltic coasts—and also in the Low Countries and Scandinavia. These regions comprise the area dominated by the Hanseatic League in the Middle Ages, hence the name of the tankards. Other regions of Europe were evolving their own special types of vessels for beer and wine, which, with a few modifications, remained standard for centuries. Thus, it is a very simple matter to distinguish between baluster jugs from London and *pichets* from Paris or between wine flagons from Switzerland and those made in the Low Countries, Burgundy, the Main regions of Franconia, southern Germany, and the Rhineland. The type of a baluster jug made in the region around Frankfurt-am-Oder and in Brandenburg in northeastern Germany is particularly elegant and distinguished looking. The few jugs of this type that have survived date from about 1500.

In all of the districts bordering the Rhine, vessels with flat lenticular (the shape of a double-convex lens) bodies are relatively common. They were used as canteens—sometimes as tankards, in which case they had a base that acted as a stand.

16th century to modern. The Baroque era saw the production of many different types of drinking and pouring vessels, often made of pewter. The guilds, for instance, commissioned drinking vessels in the shape of larger than life-size versions of the tools of their trade or their coats of arms. Another type of vessel was called the Welcome,

a drinking vessel that was handed around as a form of greeting or when a toast was being drunk. The body of these vessels was generally cylindrical or potbellied, with a lid and a short shaft set on a circular base.

Far fewer plain everyday plates have survived from the 15th and 16th centuries than drinking vessels and containers of the same period. The earliest pewter plates and bowls to have survived in any quantity date from the 17th century.

In the last half of the 16th century two places in Europe evolved quite independently, though simultaneously, a new technique for casting pewter. The product was a type of relief-decorated ware known as "display pewter" (*Edelzinn*), and it gave a new and brilliant impetus to the trade. The first examples were made between 1560 and 1570, and the main centres of production were Nürnberg and Lyon. In the beginning the technique used was not the same in both towns. Whereas in France, relief pewter was cast in engraved brass molds worked with a burin, in Nürnberg etched molds were used. This suggests that the two towns were not influenced by each other in any way. Later on, however, Nürnberg pewterers were strongly influenced by the work of a celebrated French pewterer, François Briot, who was active in Montbeliard, in the county of Württemberg.

"Display pewter"

The first master pewterer documented to have made relief pieces in Lyon is Roland Greffet, between 1528 and 1568. One can assume that it was he who invented this type of work. A school producing tankards and dishes with relief decoration soon grew up in Lyon. The most common decorative motif was an arabesque, which was used in a variety of ways and can be thought of as the leitmotif for the work of this group of artists. The master of relief pewter was François Briot. His most famous piece is the Temperantia Dish (Figure 168), which takes its name from the allegorical figure of Temperance or *Temperantia* that appears in the centre of it. It dates from 1585–90.

Giraudon—Art Resource



Figure 168: Temperantia Dish, relief-decorated "display pewter," by François Briot, 16th century. In the Louvre, Paris. Diameter 45 cm.

Pewter with etched relief decoration was made by Nürnberg pewterers from the last third of the 16th century onward. The earliest piece made by Nicholas Horchhaimer, bearing the date 1567, is a dish cast in an etched mold with an allegorical figure representing Fame, or *Fama,* in the centre and historical scenes or incidents from classical mythology around the edge. Other large dishes made by Horchhaimer and his contemporary Albrecht Preissensin are again decorated with themes from classical antiquity or sometimes with biblical scenes; for smaller plates they kept to abstract decoration.

The use of etched molds did not remain fashionable in Nürnberg for long, and toward the end of the 16th century engraved molds were being used here as well. The work of François Briot was copied by Caspar Enderlein, who modelled his own Temperantia Dish directly on Briot's. The decoration on the ewer that went with it was modelled on Briot's Mars Dish and on a piece known as the Suzannah Dish, which is also attributed to Briot.

In the second quarter of the 17th century, smaller relief plates superseded the big dishes and jugs made in Nürnberg. The Mannerist allegories that had been in favour completely disappeared, to be replaced by scenes from the Old and New Testaments, equestrian portraits of the German emperors with the electors round the edge, and luxuriant floral decorations. These plates are no more than about seven inches (18 centimetres) in diameter and are generally flat and disk-shaped. The molds were no longer made by the pewterers themselves but by professional mold cutters, who occasionally added their own monograms. Since molds were often sold by one workshop to another and then to another, one sometimes finds plates cast in the same mold but with different touches. Small decorative plates of this type were so popular that they continued to be made as late as the 18th century. There are no less than nine different models for a plate with an equestrian portrait of Ferdinand III of the House of Habsburg, who was crowned emperor of Germany in Nürnberg in 1637. Similar plates depicting Gustavus Adolphus of Sweden, the Emperor of Turkey, and Duke Eberhard im Bart of Württemberg were also produced.

Few places, apart from Nürnberg and France, had a flourishing trade in relief pewter. A few master pewterers in Saxony did execute relief decoration, however, mainly on jugs; they adapted their motifs from lead or bronze plaquettes made in southern Germany. Plates bearing the arms of Switzerland were also produced by Swiss pewterers in the 17th century. They have scenes taken from the history of Switzerland. The golden age of relief pewter, which had begun about 1570, ended in the third quarter of the 17th century. During this period, individual craftsmen had elevated pewter from its humble status as a material from which functional articles were made to one in which brilliant artistic feats could be performed. Relief pewter pieces were solely works of art, nonfunctional objects valued as showpieces.

Pewter dishes made in Italy in the 16th and 17th centuries have chased, etched, engraved, or chiselled decoration and lean heavily on artists working in brass or bronze for their designs. An independent pewter trade does not seem to have existed in Italy on anything like a large scale until the 18th century.

Domestic pewter in northern Europe

After the Thirty Years' War the production of functional articles in pewter noticeably increased in northern Europe. Besides a very large number of different types of jugs, each region specializing in its own characteristic design, there were plates and dishes used at table and also basins and bowls, drinking mugs, and screw-top flasks.

Yet pewter was already feeling the draught of competition by the end of the 17th century. In this time pewter began to be superseded by products of other branches of the decorative arts. Its first rival, faience ware, was initially no more than an inferior substitute for porcelain; but because the factories that were soon springing up everywhere were able to produce very large quantities of faience, they inflicted heavy damage on the pewter trade. Faced with this situation, the pewterers switched to imitating the designs used by the silversmiths, in the hope of gaining favor in the more ambitious middle class circles. This attempt was successful; and, from the first quarter of the 18th century onward, "silver-type pewter" gained a firm hold, soon influencing the production and appearance of pewter ware made in the Regency and Rococo periods.

By about the middle of the 18th century, an ever-widening variety of articles was being made: the pewterers were able to supply anything from a spoon to a whole dinner service, including mustard pots, sauceboats, and spoons for serving punch. But this period of prosperity

was short-lived. By the third quarter of the 18th century, pewter was rivalled both by porcelain, which could now be produced relatively cheaply by several factories in Europe, and by the even cheaper English earthenware that flooded markets on the Continent. This new development sealed the fate of the pewter trade. Towns that once had 20 or 30 busy and successful workshops had no more than one or two by the beginning of the 19th century.

Although in Germany the demand for pewter seems to have increased for a few years after the Napoleonic era, particularly in country districts, by the middle of the 19th century industrialization finally put an end to a trade that had flourished for centuries.

In the second half of the century, when stylistic imitations were all the rage, pewter vessels were produced in the Neo-Baroque, Neo-Rococo, Neo-Gothic, Neo-Renaissance, and other styles that followed the many historicizing trends that emerged. Yet these pieces were made more often by mechanized metalworking factories than by pewterers. The Art Nouveau style that became fashionable at the end of the 19th century brought about a revival of pewter production; and individual firms succeeded in making original, well-designed pieces that are often of considerable aesthetic importance. The firm of Kayser in Oppum near Krefeld played a leading part in this revival (Figure 169).

By courtesy of the Museum fur Kunsthandwerk, Frankfurt am Main

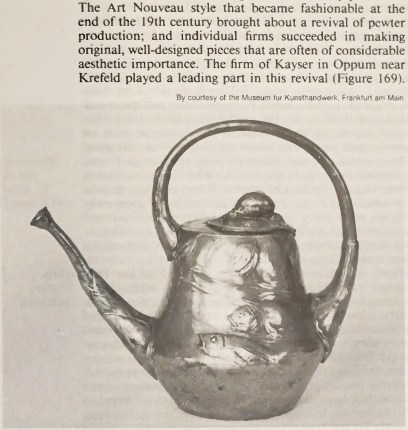


Figure 169: Pewter watering can in Art Nouveau style, by the firm Kayser, at Oppum near Krefeld, Germany, c. 1900. In the Museum für Kunsthandwerk, Frankfurt am Main. Height 21.5 cm.

But the outbreak of World War I spelled the end of Art Nouveau—whose heady run of success had anyway been short-lived—and with it the end of old pewter.

(H.-U.H.)

IRON

Ironwork is fashioned either by forging or casting. Wrought iron is the type of ironwork that is forged on an anvil. There are no fabrication similarities to cast iron, which is poured in a molten state into prepared sand molds.

Wrought iron

Wrought iron is fibrous in structure and light gray in colour. It can be hammered, twisted, or stretched when hot or cold. The more it is hammered, the more brittle and hard it becomes; but it can be brought back to its original state by annealing (heating and then cooling slowly). It will not shatter when dropped.

From earliest times, the smith has had a forge to heat the iron, an adjacent water tank in which to cool it, an anvil on which to form it, in addition to a wide assortment of hammers and tools. The most important tool is the anvil. The English type, generally used for forging wrought iron, has a flat top surface, which is used as a solid base for hammering the heated iron into shape, for welding, for splitting, or for incising decorative chisel marks in the hot iron. One end of the anvil is shaped like a pointed cone and is used for forming curved surfaces. The other blunt end, or heel, has one or two square or rectangular holes on top, into which fit various tools. From the anvil is derived the expression "to strike while the iron is hot," and this implies spontaneity and rapid hammer blows. The wrought-iron craftsman should not be expected to repeat with meticulous exactitude one intricate component after another. In fact, wrought iron by a master craftsman is esteemed for the variations that naturally occur.

Components of wrought-iron design

The individual components of a wrought-iron design are often plain or twisted rods, with or without chisel-mark incisions. They are frequently composed as a series of straight, parallel members or in combination with scrolls, or as a repeat design of some geometric shape such as the quatrefoil. Where two curved members are tangent, they are characteristically secured together by bands or collars, rather than by welding. Where two straight bars intersect, it is accredited craftsmanship to make the vertical bar pierce or thread the horizontal member. Grilles consisting of two series of parallel small-diameter rods, one series at right angles to the other, were sometimes interlaced or woven.

Depending upon the depth of the relief, various fabrication techniques may be employed for repoussé, or three-dimensional, ornamental wrought ironwork. Sheets $^1/_{16}$ inch (1.6 millimetres) or less in thickness generally are used. The general configuration of the modelling is obtained by beating the back of the sheet; the final details are embossed on the front face. The finer the scale and detail, the more work must be done when the iron is cold. A repoussé design may be pierced; but this term usually connotes a solid sheet forged into a mask, a shield, or an entire embossed panel. The traditional means of setting off a cutout repoussé design was to superimpose it on a vermillion-coloured background panel. Modern approximations of repoussé work consist of mechanically stamped designs touched up with random hammer blows.

(G.K.Ge.)

The most difficult way of decorating iron is to carve it. This involves fashioning figurative or decorative motifs out of the metal ingot with especially strengthened tools, using the material in the same way that the sculptor handles wood or stone. Only very precious iron articles are carved, such as coats of arms or pieces that are specifically designed to be displayed as works of art.

(H.-U.H.)

Cast iron

Cast iron is melted in a furnace or cupola, stoked with alternate layers of coking iron, then poured into prepared sand molds. After the cast iron cools in the mold, the sand is cleaned off, and the work is virtually complete. Its shape is fixed, and while a casting can be slightly trued up by the judicious use of a hammer, it is in no sense as workable as wrought iron. Thus, ornamental features in cast iron cannot be chased and polished as in cast bronze. If the ornamental cast-iron details are not replicas of the original pattern, the only recourse is to make a new casting. Because it is brittle, cast iron is almost certain to shatter if dropped.

Since it is cast in a mold, certain forms are more suitable to cast iron than to wrought iron. For example, if repetitive balusters, or columns, or panels with low-relief ornamentation are desired, cast iron is the most suitable material.

(G.K.Ge.)

Early history. The earliest recorded iron artifacts are some beads, dating from about 3500 BC or earlier, found at Jirzah in Egypt. They are made from meteoric iron, as are a number of other objects of only slightly later date that have been found both in Egypt and Mesopotamia. The earliest known examples of the use of smelted iron are fragments of a dagger blade in a bronze hilt, dating from the 28th century BC, found at Tall al-Asmar (modern Eshnunna), in Mesopotamia, and some pieces of iron from Tell Chagar Bazar, in the same area, of approximately the same date. There is, however, no evidence of any extensive use of iron in either Egypt or Mesopotamia before the end of the 2nd millennium BC. In Asia Minor, on the other hand, iron was probably used regularly from at least as early as 2000 BC; and it seems likely that the first true iron industry was established there in the second half of the 2nd millennium BC.

From the ancient Near East the knowledge of iron work-

ing was transmitted to Greece and the Aegean, probably at the beginning of the 1st millennium BC, whence it spread gradually to the rest of Europe. By the 6th century BC, it had been widely disseminated over central and western Europe.

Iron was at first apparently regarded as a precious, semi-magical material, presumably because of its rarity and its connection with meteorites. But once it had become common, as a result of increased knowledge of the technique of smelting ore, it seems to have been used, at least in Europe, almost exclusively for objects of utility. A few Belgic firedogs and at least one amphora, skillfully forged in iron, with decorative terminals in the form of animal heads, are known; but the practice of forging iron into decorative shapes does not seem to have become general until the Middle Ages.

A few cast-iron objects dating from classical times have been found in Europe. The extreme rarity of these, however, suggests that they were only produced experimentally. The earliest known evidence for the general use of cast iron comes from China (see below), and it does not seem to have been produced regularly in Europe before the 15th century. (C.Bl.)

Belgium and Holland. The ironwork of these two small countries prior to the 15th century was in no way inferior to that produced elsewhere. Yet so few pieces remain that the significance of craftsmen of the Low Countries has often been underestimated. During the 15th century, design and craftsmen from the Low Countries began to make their influence evident across the channel in England. Representative examples of this period are in the Hervormde Kerk at Breda; the treasury door of the cathedral at Liège; and hinges of the church of Notre Dame, at Hal. The beautiful spires of Bruges, Ghent, and Antwerp should be mentioned.

Diversified forms of ironwork

During the first half of the 16th century, before the Spanish occupation, there were diversified forms of ironwork, such as protective grilles for doors, windows, and chapels, often in fleur-de-lis patterns; window gratings of vertical bars, frequently octagonal in section; and interlacing bars, producing rectangular or lozenge-shaped patterns. Only a few examples still exist: some lunettes in the Hôtel de Ville of Brussels; a tabernacle grille from the chapel of the counts of Flanders and a window grille from the Cathedral of St. Bavon, both from Ghent (Victoria and Albert Museum); and hinges at the Hôtels de Ville of Bruges and Ypres (Flemish Ieper). Few Renaissance screens have survived.

During the second half of the 16th century, the cruelty of the Duke of Alba and his 20,000 troops, together with the threat of the Inquisition, drove hundreds of artisans to England. After the Spanish domination there was little indigenous design in Holland and Belgium, and such ironwork as was produced fell under the spell of French imports. (G.K.Ge.)

England. The initial use of wrought iron was purely protective because violent attacks were frequent, and doors had to be strengthened with massive ironwork inside and out. Window openings, especially those of the treasuries of mansions and cathedrals, were for similar reasons filled with strong interlacing bars of solid iron; a good example remains at Canterbury cathedral. When, in the course of time, the need for protective barriers ended, there was greater freedom of work and a definite trend toward ornamentation. Throughout England, medieval church doors are found with massive iron hinges, the bands worked in rich ornamental designs of scrollwork, varying from the plain hinge band, with crescent, to the most elaborate filling of the door. Examples exist at Skipwith and Stillingfleet in Yorkshire, many in the eastern counties, others in Gloucester, Somerset, and the west Midlands. The next important application of ironwork came with the erection of the great cathedrals and churches, whose shrines and treasures demanded protection. Winchester Cathedral possesses the remains of one screen with a symmetrical arrangement of scrollwork. Tombs were enclosed within railings of vertical bars with ornamental finials at intervals, such as that of the Black Prince at Canterbury. A new development appeared in the early years of the 15th century when the smith, working in cold iron, attempted to reproduce Gothic stone tracery in metal. This work was more like that of a woodworker than of a smith, often consisting of small pieces of iron chiselled and rivetted, and fixed on a background of sheet iron. Many small objects such as door knockers, handles, and escutcheons were executed in the same manner. A typical monumental example is in Henry V's chantry at Westminster Abbey; but the most magnificent is the great grille at St. George's Chapel, Windsor, made to protect the tomb of Edward IV.

Decoration for church doors

The development of the art of smithing during the Renaissance period was very uneven in the various countries of Europe. In 16th-century England the smith fell behind and seemed to have lost interest, producing no very great or important work. He continued to make iron railings, balconies, and small objects for architectural application, such as hinges, latches, locks, and weathercocks. But toward the end of the 17th century, there was a growing interest in beautifying houses and laying out gardens and squares, with a commensurate demand for balconies, staircases, and garden gates. The man to whom the credit is usually given for the revival of ironwork in England was Jean Tijou, a Frenchman who, together with many of his Protestant fellow craftsmen, had been forced to leave his country owing to the revocation of the Edict of Nantes in 1685. After some years in The Netherlands he went to England in 1689, where he enjoyed the patronage and favour of William III. His most important works for his royal patron are to be seen in the immense mass of screens and gates with which he embellished Hampton Court palace. He also executed work at Burleigh house, Stamford. Probably by the Queen's wish he was associated with the architect Sir Christopher Wren, then engaged on the rebuilding of St. Paul's Cathedral. Wren apparently did not particularly like ironwork and probably exercised some restraint on Tijou, with the result that his work at St. Paul's is more dignified and freer from appendages than that of Hampton Court.

18th-century ironwork

There is a great amount of fine ironwork of the 18th century in London in the form of gates, railings, lamp holders, door brackets, balconies, and staircases; in almost every suburb there are gates and brackets. The precincts of the colleges of Oxford and Cambridge, as well as almost every old town in England, furnish a variety of handsome work. Throughout the 18th century the smith was a busy man; the general tendency of his work, unaffected by the Rococo movement on the Continent, was toward a less ornate but more characteristically English style—perpendicular, severe, lofty, and commanding, as contrasted with Tijou's French love of richness and mass of details.

At the end of the 18th century the work of the architect brothers Adam shows a departure from true smithing; its slender delicate bars are enriched with rosettes, anthemia, and other ornament in brass or lead. The effect is pleasing and harmonizes with the architecture with which it is incorporated.

During the first half of the 19th century, the art of the smith was largely eclipsed by that of the iron caster. But under the stimulus of the Victorian Gothic revival and later of the Art Nouveau movement, there was a renewal of interest in the decorative use of wrought iron, and much excellent work was produced.

Medieval door-hinge ornaments

France. Medieval door-hinge ornaments were not basically different from those in England; and beautiful work is found on church doors, especially in central and northern France. It reaches a height of greater elaboration and magnificence than in England, the culminating example being the west doors of Notre Dame, Paris, the ironwork of which is so wonderful that it was attributed to superhuman workmanship. Grilles at Troyes and Rouen also reveal a high standard of excellence. Working the iron cold and employing methods associated with carpentry was immensely popular; it was applied to small objects such as door handles, knockers, and above all to locks, which exhibit an amazing amount of detail and a remarkable delicacy of finish.

The Gothic tradition survived in France until well into the 16th century and was marked by the production of work of the highest skill, largely in the form of locks,

knockers, and caskets of chiselled iron. The introduction of the Renaissance style did not radically alter the direction of the smith's art—a strange fact when it is remembered that Germany and Spain were fabricating works of enormous size and magnificence in wrought iron. France, like England at that time, was content to make door furniture, in the form of locks, keys, bolts, escutcheons, and the like, but did little ironwork of any great size. A school of locksmiths came into being under Francis I and Henry II, working from designs by Androuet du Cerceau in the 16th century and those by Mathurin Jousse and Antoine Jacquard in the 17th. The bows (a loop forming the handle) and wards (notches) of keys were of unusually intricate design and the locks of corresponding richness. Representative pieces may be seen at the Victoria and Albert Museum. Among them is the famous Strozzi key, said to have been made for the apartments of Henry III, the bow of which takes the favoured form of two grotesque figures back to back. But as far as architectural ironwork was concerned, France remained almost at a standstill until the accession of Louis XIII in 1610. Under that monarch, a worker at the forge himself, came a great revival, which, by the end of the 17th century, had attained a marvellous pitch of perfection. It proved to be the beginning of a new movement, the force of which made itself felt in the adjoining countries and inspired ironworkers with new energy. From the accession of Louis XIV, the French ironworkers must be acknowledged as the cleverest in Europe, combining as they did good and fitting design with masterly execution. Their designs were often very daring, exploiting all the latent and previously unexplored possibilities of iron. They recognized its great adaptability and took every advantage of it, at the same time being conscious of its limitations. Their forms of expression were endless.

Ironwork in the reign of Louis XIV

Screens and gates were needed for parks, gardens, and avenues, staircases for mansions and palaces, screens for churches and cathedrals. Among celebrated designers were Jean Lepautre, Daniel Marot, and Jean Berain. Earlier work had been of a simple character—balconies, for instance, being in the form of a succession of balusters—but as the smith became more versatile and imaginative, they took the form of panels of flowing curved scrolls, rendered with a freedom never attained before, while constructive strength was observed and symmetry maintained. Enrichments were usually attached in hammered sheet iron. These may be considered the distinguishing features of Louis XIV work, such as that at St. Cloud, Chantilly, Fontainebleau, and elsewhere. But under Louis XIV all previous efforts were surpassed in the work for his palace at Versailles.

Rococo design

The art of ironwork received a further impetus by the introduction of the Rococo style. The movement, initiated in 1723, was due principally to the imagination of two artists, Just-Aurèle Meissonier, architect, and Gilles-Marie Oppenordt. There was a balanced asymmetry in the design and fantastic curves with a luxury of applied ornamentation. To the French smith it furnished the opportunity for a yet greater display of his skill. He was clever enough to secure a feeling of stability in his work by counterbalancing swirling masses of ornament with straight constructional lines; he knew how to introduce an iron screen of Rococo style into a Gothic church or cathedral without giving offense to the eye or arousing any uncomfortable feeling of incongruity.

Later in the 18th century, ironwork took on a more classical appearance as a result of the general revival of interest in ancient art; and many Greek and Roman details were introduced into the ornamentation. The amount of work executed was prodigious, and its beauty and craftsmanship may be seen in most cities of France. Nearly all of the adjacent countries, with the exception of England, were seized with the desire to imitate the French Rococo style.

Germany. In the Romanesque period in Germany, bronze was preferred to iron; the earliest examples of ironwork are thus later than those of France and England.

German door hinges

The first iron grilles were imitations of French work, with C-scrolls filling spaces between vertical bars. Typical examples of door hinges prior to the 14th century were those at Kaisheim, St. Magnus Church, Brunswick, and St. Elizabeth's Church, Marburg (the latter having a curious cross in the middle). Throughout the Gothic period in Germany, the imitation of natural foliage was the basis of design.

There were no new marked developments in ironwork during the 14th century. Smiths confined their efforts mostly to hinges. Until this period the vine had been the only motif for elaborate hinges; but flat, lozenge-shaped leaves were introduced, such as those at Schloss Lahneck on the Rhine.

During the 15th century, grilles became more popular. One of the best examples is the grille in the Monument of Bishop Ernst of Bavaria, Magdeburg cathedral (c. 1495), with elaborate Gothic tracery, nine columns, and a cornice. In hinges the cinquefoil displaced the quatrefoil, as at Orb, Oppenheim, and Magdeburg. The Erfurt cathedral was enriched with notable hinges having the vine pattern interpolated with rosettes and escutcheons of arms. Hinges for houses usually were the plain strap type, but when ornamented they consisted of superimposed layers of sheet iron. As in other parts of Europe at this time, pierced sheet iron was fashioned into tracery of a semi-architectural nature, much like Gothic windows. Pierced ornament and twisted rods were often combined to form grilles, with their extremities beaten into complicated foliage forms.

During the Renaissance, ironwork in Germany was in use everywhere and for every purpose: for screens in churches, window grilles, stove guards, gates, fountain railings, well heads, grave crosses, door knockers, handles, locks, iron signs, and small objects for domestic use. Smiths were their own designers and more often than not planned intricate devices merely to show their skill in executing them. They set no limits to their problems; and so far as manipulative excellence went, the German smiths were the foremost in Europe. But clever as their workmanship undoubtedly was, their designs frequently showed a lack of stability and a tendency to run riot. Thus, many of their most imposing works consist largely of filling panels with elaborate, interlacing scrollwork, and the sense of constructional and protective strength is missing.

An abundance of smiths' work is to be found in the

Bildarchiv Foto Marburg

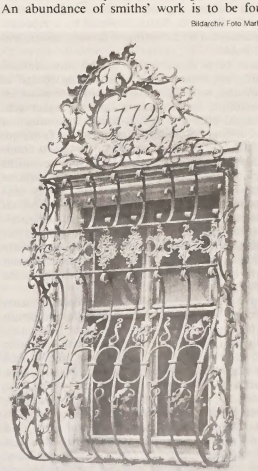


Figure 170: Rococo style wrought-iron window grille from a house on Winklerstrasse, Nürnberg, 1772.

Southern German work

southern parts of Germany. Iron bars, circular in section, were most frequently used; and the most common features are interlacing bars and terminations of flowers with petals and twisted centres, foliage, or human heads. All of these characteristics occur with almost monotonous repetition, witnessing to skill but also to lack of imagination and sense of design. The style may be studied in many German and Austrian cities, such as Augsburg, Nürnberg, Frankfurt, Salzburg, Munich, and Innsbruck (Figure 170).

The German smith gave much attention to door knockers and handles, enclosing them in pierced and embossed escutcheons, and devised locks with very involved mechanism. German influence made itself strongly felt in Switzerland, Austria, and Czechoslovakia.

The Baroque and Rococo periods are distinguished by a perfection of detail that exceeded that of German Medieval or Renaissance ironwork. Smiths used wrought iron as though it were a plastic material, meant to be employed in extravagant forms wherever possible. Some examples are at Zwiefalten, Weingarten, and Klosterneuburg. In the late 18th and early 19th centuries, cast ironwork of outstanding quality was produced in Germany, notably at the Prussian royal foundry established in 1804.

Italy. The few extant examples of ironwork in Italy prior to the 14th century indicate a wide appreciation of how the material could best be worked with only the tools of the smith. Some noteworthy examples are the chancel grille at the left of the nave, Orvieto Cathedral (1337); the grille around the Scaligeri tombs of Verona (*c.* 1340); the grille at the baptistery of Prato cathedral (1348); the chancel screen in the sacristy chapel of Sta. Croce, Florence (1371); and the grille to the Capella degli Spagnoli, Sta. Maria Novello, Florence.

Italian ironwork before the 16th century

Until the 16th century, Italian smiths respected the natural characteristics of wrought iron by relying almost entirely upon those forms that could be wrought with hammer and anvil. The grille was usually made by dividing it into regular panels with vertical and horizontal bars (sometimes triangular in section and enriched with dentils, or small, projecting triangular blocks). Often the quatrefoil filled some or all of these panels; they were made in Tuscany from a pierced plate and in Venice from separate scrolls collared together. A noted example is in the Palazzo della Signoria, Siena, crowned by a repoussé frieze and surmounted by a cresting of flowers, spikes, and some animal heads.

It might have been thought that in the fountainhead of the Renaissance, ironwork would have proceeded at the same pace and with the same brilliant success as architecture, sculpture, bronze casting, and the other arts. Strangely enough, little use of it is found in connection with the fine buildings of the revival. Bronze was favoured; and what in other countries is found in iron has its counterpart in Italy in bronze. As time went on the smiths grew less inclined toward the more difficult processes of hammering and welding and contented themselves ultimately with thin ribbon iron, the various parts of which were fastened together by collars. Work of the later periods may be distinguished, apart from the design, by this feature, whereas the English and French smiths vigorously faced the hardest methods of work, and the German and Spanish smiths invented difficulties for the sheer pleasure of overcoming them.

Notable centres of artistic ironwork were Florence, Siena, Vicenza, Venice, Lucca, and Rome, where important pieces may be found in the form of gates, balconies, screens, fanlights (semicircular windows with radiating sash bars like the ribs of a fan), well covers, and a mass of objects for domestic use, such as bowl stands, brackets, and candlesticks.

Screenwork

In screenwork the favourite motif was the quatrefoil, which has been found with many variations ever since the 14th century. Early examples are strong and virile, but later ones tend to weakness. The C-shaped scroll is also used in many combinations. The churches and palaces of Venice contain many examples of these popular designs. Peculiar to Italy are the lanterns and banner holders such as may still be seen at Florence, Siena, and elsewhere, and the rare gondola prows of Venice. Of the ironworkers of the early Renaissance, the most famous was the late-15th-century craftsman Niccolo Grosso of Florence, nicknamed "Il Caparra" because he gave no credit but insisted on money on account. From his hand is the well-known lantern on the Palazzo Strozzi in Florence, repeated with variations elsewhere in the same city. Siena has lanterns and banner holders attached to the facades of its palaces, and lanterns are still to be seen at Lucca and a few other towns.

The decadence of 17th- and 18th-century ironwork paralleled that of architecture. Designs were borrowed directly from France and Germany. The metal was too often worked cold, using thin members; and the resulting construction was flimsy. Scrolls were often encased in thin, grasslike leaves. Conventional or naturalistic flowers were tacked on as seeming afterthoughts. Instead of using rods and bars, ribbonlike bands were used, with cast ornaments pinned on. Intersecting tracery was copied from Germany. The best examples of this period are confined to Venice and northern Italy, such as the screen in the south aisle chapel of S. Ambrogio, Milan; the chapel enclosure in S. Pietro, Mantua; and the screen in the Palazzo Capodilista, Padua.

Spain. Prior to the 15th century, Spanish ironwork was basically similar to that in France and England. The Spanish smith accepted the limitations imposed by anvil and ancillary tools; but he skillfully exploited to the limit all manner of variations—twisting square rods, coiling flat bars into C-shaped scrolls of all sizes, and devising imaginative crestings to surmount the top of church chapel screens or domestic window grilles. Many Moorish craftsmen of extraordinary ability were enticed to remain in Spain as the Moors were slowly pushed southward; the resultant blending of Gothic with Moorish resulted in the Mudejar style.

Ironwork of the Renaissance period from about 1450 to 1525 reached a height of grandeur and magnificence attained in no other country. Of all the Spanish craftsmen the smiths were the busiest, especially during the 16th century. The ironwork products that for more than a century dominated the craft are the monumental screens (*rejas*) found in all the great cathedrals of Spain. These immense structures, rising 25 to 30 feet (7.5 to nine metres) show several horizontal bands, or tiers, of balusters, sometimes divided vertically by columns of hammered work and horizontally by friezes of hammered arabesque ornament. Usually such screens are surmounted by a cresting, which is sometimes of simple ornament but more often a very elaborate design into which are introduced a large number of human figures. Shields of arms are freely incorporated; and the use of bright colour, silvering, and gilding adds to their impressive beauty. The great balusters were always forged from the solid, and their presence in hundreds demonstrates the extraordinary skill and power of the Spanish smith. In many cathedrals two of these monumental *rejas* are found facing one another. There is at least one in every large cathedral—Barcelona, Saragossa, Toledo, Seville, Burgos, Granada, Córdoba, and many others.

Monumental *rejas*

Ironwork on a smaller scale is found in gates, balconies, and window screens; wrought-iron pulpits also exist. Panels of hammered and pierced iron, heightened with colours and gilding, were used in connection with domestic architecture; and many doors were ornamented with elaborate nailheads or embossed studs.

(W.W.W./G.K.Ge.)

United States. The characteristics of the earliest ironwork in the various colonies naturally reflected those of the parent countries. The English were more sparing in its use in the New England Colonies than were the Germans in Pennsylvania or the French in Louisiana. In the 17th and 18th centuries ironwork was used mostly for such practical purposes as weather vanes, foot scrapers, strap hinges, latches, locks, and particularly for the necessities and conveniences for fireplaces (firedogs, cranes, skewers, toasters, kettle warmers, and spits). It was not until the late 18th century, when the threat of Indian raids and food shortages had waned and the established communities enjoyed a sense of tranquillity and prosperity, that smiths fashioned wrought iron into railings, fences, grilles,

Utilitarian ironwork

gates, and balconies. Square or flat iron bars were generally used to produce designs that were usually light, airy, and graceful and rather in contrast to the contemporary European preference for sturdier forms.

Gradually, ironwork designs tended to develop characteristics of an American or composite nature, as a logical consequence of the diverse origins of colonists and smiths. An innovation that appeared toward the end of the 18th century was the combination of structural wrought-iron rods or bars with lead or cast-iron ornamental features. While the use of wrought iron declined in the 19th century, during its last quarter the use of cast-iron columns and panels for nonresidential buildings increased. These designs, timid or bold, decorative or structural, engendered the prototypes of commercial buildings for the ensuing decades.

Because the life of structures in U.S. cities has been short, there are few examples of 18th- or early 19th-century ironwork extant in New York City, not many more in Boston, some in Philadelphia, but more in and near Washington, D.C., such as the excellent balconies and railings at the Octagon (headquarters of the American Institute of Architects). Charleston, South Carolina, has a rich legacy in gates, notably those at numbers 12, 23, and 36 Legare Street, 63 Meeting Street, and an unusually beautiful pair at St. Michael's Church.

New Orleans ironwork

New Orleans has more ironwork than other U.S. cities, thanks to a group of citizens dedicated to the preservation of the old French Quarter. Its earliest ironwork was forged by Spanish and French smiths. Unfortunately, fires, rust, and remodelling have so taken their toll of the Spanish ironwork that almost the only remaining example of importance is the gateway of the Cabildo (town hall). It has moldings beaten from solid bars, like many of the old *rejas* in Spanish cathedrals. After the Louisiana Purchase in 1803, the influx of ironworkers from northern states brought about a broadening of influences that is apparent in designs and techniques. Ironwork of New Orleans can be roughly divided into three periods: (1) forged wrought iron by French and Spanish artisans with strongly marked European characteristics; (2) a transitional period with wrought-iron structural members embellished with cast-iron ornaments in the Directoire and Empire styles of France, plus some U.S. innovations; and (3) entire grilles, screens, and trellises made entirely of cast iron. No other city in the U.S. has two- and even three-story iron porches and balconies that can compare with those of New Orleans. Some of these lacy structures, such as those on St. Peter Street (Figure 171), were built above the sidewalks. Balconies sometimes not only extended across an entire facade but continued around a corner.

By courtesy of the Historic New Orleans Collection



Figure 171. Wrought- and cast-iron balconies along St. Peter Street, in the Vieux Carré, New Orleans, c. 1838–40.

Mid-19th century onward. Distinctive national characteristics in the design of ironwork gradually tended to disappear in Europe because of increased travel and communications between countries. The influence of French Renaissance architecture (modified or revived) continued to exert a viable effect where the acceptance of the Art Nouveau (last quarter of the 19th century) was flaccid or denied. In England, however, 18th-century designs continued with slight modifications. In the U.S. probably the most important force, prior to World War I, was exercised by architects trained in Paris, with the result that ironwork designs were similar to French work of this period.

The increased mechanization of all forms of manufacture understandably affected the character and use of ironwork. As the cost of cast iron came down, its use increased. Because wrought iron is produced by hand by beating red-hot iron on an anvil, not much change was possible through increased mechanization, whereas the casting of molten iron lent itself to improved equipment and techniques. The lowered cost of duplicating ornamental cast-iron components and the introduction of structural steel parts expanded the usage of ironwork to the modest building, whereas it had been generally confined to public or monumental structures. Foundries in the U.S. established a flourishing business in pierced cast-iron panels, modelled after Louisiana porch trellises.

Effects of increased mechanization

Compared with prior periods, the last half of the 19th century will scarcely be commemorated as introducing enduring or beautiful ironwork forms. It was not until the first quarter of the 20th century that a master craftsman-designer gave impetus to a new conception of design forms and textures. Edgar Brandt of Paris broadened the scope of decorative usage by the rich inventiveness of his compositions and by an entirely original approach that resulted in a wrought-iron texture that is akin to beaten silver. Examples of his work at the Exposition des Arts Décoratifs Modernes at Paris in 1925 had an immediate effect upon ironwork designed and executed in the U.S. during the great building boom that lasted until about 1930. During this period, both wrought and cast iron enjoyed an unprecedented period of popularity not only in the form of bank screens, entrance doors, and grilles in public buildings but as decorative grilles and gates in private homes. In many cases the craftsmanship equalled that of representative examples of the Gothic or Renaissance periods in Europe.

One of the most gifted and dedicated iron craftsmen in the U.S., Samuel Yellin of Philadelphia, raised the standards of wrought-iron craftsmanship to its apex during the 1920s. He not only trained an atelier of craftsmen for the first time in the U.S., but by his efforts wrought iron was recognized as capable of enriching even the most monumental building. Yellin's influence, however, was ended by the Depression of the early 1930s. As building activity declined after 1930, so did the use of ironwork; and it did not increase with the revival of building after World War II. (G.K.Ge.)

LEAD

Lead has two main uses in which some artistic purpose may be served: in architecture, as a material for roof coverings, gutters, piping, and cisterns; and in decorative art, as a material for sculpture and applied ornament. As an architectural material it has the advantage of being easily worked and yet offers great resistance to climatic conditions. The low melting point of lead and its relative freedom from contraction when solidifying make it particularly suitable for casting, and it has been used as a substitute for bronze or precious metals.

Antiquity. The earliest known lead sculptures are small

Earliest known lead sculptures

votive figures found at Troy and Mycenae. In the Hellenistic period lead sarcophagi were known, and the Romans made much use of the metal. Large amounts of worked lead in various forms have been found in those parts of England where the Romans had permanent settlements.

Middle Ages. England was one of the main lead-producing areas in the Middle Ages, and lead was more widely employed there than on the continent of Europe. In the 12th century the German monk Theophilus, in his treatise on metalworking, refers to lead only in connection with casting rods for stained-glass windows and as a material through which silver sheets might be hammered; but in England at about the same time a remarkable series of lead fonts was cast, of which 16 still survive in position, the most famous being those at Walton-on-the-Hill, Surrey, and at Wareham and Dorchester in Dorset. Lead was also used in the Middle Ages for church roofing; and it was used, doubtless because of its cheapness, for the small badges or medallions sold to pilgrims at the great medieval shrines. Lead could even be useful, in the proper disguise, to simulate rich ecclesiastical objects, for not all religious institutions were wealthy: a group of 14th-century caskets covered with lead tracery, gilded to look like precious metal, have survived in church treasuries. These were used as reliquaries, but some were originally made for secular purposes.

Renaissance to modern. The Renaissance passion for collecting bronze medals and plaquettes led to a demand for cheap replicas, and these were made with great precision in lead. The metal also played an important role in the goldsmiths' trade. The fashion for elaborate relief ornament of the Renaissance and Mannerist periods called for a degree of skill in modelling that was beyond the powers of the average goldsmith. The practice therefore grew up for the pattern makers of Augsburg and Nürnberg, Germany, to sell lead models of ornamental details and figures from which goldsmiths working elsewhere could in turn make molds. An extensive collection of these models is preserved in the Historisches Museum, Basel, Switzerland. The trade expanded to include large medallions and plaquettes, the chief masters of which were the German goldsmiths Peter Flötner, Jonas Silber, and the Master H.G. (Hans Jamnitzer) and the Dutch goldsmith family of van Vianen. Lead in sculpture is more suitable for the production of small figures than life-size statues, which, if unsupported, become distorted through their own weight. Among the few life-size equestrian lead statues is one of Frederick Louis, prince of Wales, in the grounds of Hartwell House, Buckinghamshire, England. From the 16th century, lead appeared in England in the form of gutters and pipe heads (which carried rainwater down from the gutters), often with cast ornament. Some of the late-17th- and early-18th-century pipe heads, cast with the arms of the owner of the house and the date of erection, are important decorative features.

Lead models

An extension of the use of lead took place with the introduction of lead garden sculpture—figures, vases, and urns—in the late 17th century. An example of this work is a pair of garden vases 15 feet high at Schloss Scheissheim in Bavaria. The silvery gray colour of such sculpture and its resistance to the weather made it suitable for use in the many formal gardens that were created at this time. English garden sculpture rarely achieves any particular aesthetic status; but in 18th-century Germany and Austria lead was used for more serious sculpture by a group of artists of high standing. In the 19th century, lead was out of favour with sculptors, partly because improved transport made it possible to bring marble from Italy at low cost. Its soft colouring and the fact that it does not reflect light give it advantages, however, and it has been used in the 20th century by Aristide Maillol and by Sir Jacob Epstein, who executed the lead figure of the Virgin and Child in Cavendish Square, London. (J.F.Ha.)

Non-Western metalwork

SOUTH ASIA

Iron. The manufacture of iron by primitive smallscale methods has survived in southern India and Ceylon to the present day. The slag heaps of ancient furnaces are common, and the processes have probably been in use for more than 2,000 years; but it is unknown whether they are of indigenous invention or acquired. In southern India iron immediately succeeded stone as a material for tools and weapons, and prehistoric iron weapons began to come into use about 500 BC. The wrought-iron pillar of Delhi, set up about AD 400 by Kumāra Gupta I in honour of his father, is over 23 feet (seven metres) in height and weighs more than six tons. It demonstrates the abilities of Indian metalworkers in handling large masses of material, for not until the latter part of the 19th century could anything of the same kind have been made in Europe. There are other large iron pillars at Dhār and at Mt. Abu. (Ed.)

Gold and silver. In India, gold jewelry has been found from the Indus culture. Excavations at Takshasila (Taxila) have revealed gold and silver drinking vessels and jewelry of Hellenistic types dating back to about the 1st century AD. From the same time is the important Buddhist gold reliquary from Bimaran, Afghanistan, set in rubies and decorated with embossed figures in Gandhāra style (British Museum).

During the Gupta period (AD 320–647), vessels of Hellenistic and Persian shapes were evidently made, for they are represented in the sculpture and frescoes of the period. More Indian in style are a silver dish of the 3rd or 4th century, decorated with a Bacchanalian scene of a *yakṣa* drinking (Figure 172), and a silver bowl of the 7th century



Figure 172: Indian style embossed and chased silver dish showing a *yaksa* drinking, Kushan, found at Buddaghara, near Tank, Dera Ismāʾil Khān district, Pakistan, 3rd or 4th century AD, Gupta period. In the British Museum. Diameter 25.15 cm.
By courtesy of the trustees of the British Museum

from northern India, which is embellished with medallions in low relief (both in the British Museum). Jewelry played a very important role, and, although no original pieces have survived, it can be studied in frescoes at Ajanta and on contemporary sculptures.

In spite of the fact that gold and silver vessels have been common in India since classical times, there is very little material extant before the 17th century, when all kinds of vessels were produced in bronze, brass, copper, and, for the royal houses, in silver. Shapes and decorations vary in different regions. Delhi was famous for its craftsmen, especially in the time of Akbar in the 16th century and Jahāngīr and Shāh Jahān in the 17th. Much work was done in precious metal, and vessels and ornaments of jade were inlaid with gold and gems. Northern India is famous for its enamels. Enamellers from Lahore were brought to Jaipur in the 16th century by Mān Singh, and enamel was employed extensively in combination with goldwork and silverwork in the 17th and 18th centuries there and elsewhere. The Punjab, Lucknow, and the districts of Chānda and Cutch in Gujarāt state were long celebrated for their metalworkers. In the south, silverwork in *svamin*-style is characterized by religious-figure scenes in relief, executed in three different techniques. Craftsmen

17th-century vessels

in Tirupati put silver sheet on copper; Madras, Bangalore, and Tiruchirāppalli are known for hammered vessels with traced decoration; and Thanjāvūr (Tanjore) produced a more Baroque effect with inlays of silver in copper. From the former Travancore state, Mysore, and Bijaipur in the southwest come chased vessels with floral patterns, the lotus predominating. In the north the Hindu style is well represented by works from Vārānasi (Benares).

Persian-Islāmic influence

Persian-Islāmic influence is found in several vessel shapes; for example, ewers and basins for water and smoking furniture, such as hookas, which also have Islāmic patterns. Jewelry from the later periods employs precious stones, pearls, gold, and silver in great variety. The old types are repeated, with symmetrical arrangements of rosettes and leaves for bracelets, necklaces, pendants, rings, and foot ornaments. Very fine work in silver filigree was executed at Cuttack in Orissa and was used on jewelry and various larger items.

CENTRAL AND SOUTHEAST ASIA

Indian styles and techniques spread to the neighbouring countries. In Nepal precious metals were used in architecture; pagodas, temples, and palaces sometimes had facades richly decorated with ornaments embossed in gilt copper with settings of precious stones.

In Tibet copper and brass were usually used for vessels, but these metals were often decorated with applied silver or gold ornaments; and in eastern Tibet, especially, teapots were made of silver with gilt appliqué. While many of the ornaments are Chinese, Buddhist shapes and patterns of Indian origin were used for ritual vessels. Other ritual objects were sometimes made of silver or, more rarely, of gold, though bronze is again the common material. Silver is used for amulets and jewelry with rich settings of turquoises, carnelian, and lapis lazuli.

In Thailand, Buddhist vessels were made out of chased silver, very often in the shape of a lotus flower whose petals are decorated with other, embossed, floral and figure motifs.

Burmese gold and silver vessels

Burma is known for its chased silver vessels heavily decorated with figures and floral patterns in relief, related to the south Indian *svamin* work. The use of gold and silver vessels for domestic purposes was denied to all but those of royal blood. Good examples of earlier golden regalia are in the Victoria and Albert Museum.

In Vietnam, goldwork and silverwork of the Cham culture are preserved from the 10th century. It is exemplified by a crown and heavy jewelry made for a life-size statue found in the ruin of a temple at Mison. From later times there is a royal treasure with four crowns, various amulets, arm rings, and table services of gold, richly decorated with embossing and openwork. (B.V.Gy.)

EAST ASIA

China. *Bronze.* Bronzes have been cast in China for about 3,700 years. Most bronzes of about 1500–300 BC, roughly the Bronze Age in China, may be described as ritual vessels intended for the worship of ancestors, who are often named in inscriptions on the bronzes. Many were specially cast to commemorate important events in the lives of their possessors. The vessels were also meant to serve as heirlooms, and the inscriptions often end with the admonishment "Let sons and grandsons for a myriad years cherish and use." These ritual vessels of ancient China include some of the loveliest objects ever made by man, and as a group they represent possibly the most remarkable achievement in the whole history of metalcraft before modern times. Since the vessels can be considered sculpture, they are discussed in EAST ASIAN ARTS.

Among other ritual bronzes, bells constitute an important group. Perhaps the oldest class is a small clappered bell called *ling*, but the best known is certainly the suspended, clapperless bell, *chung*. *Chung* were cast in sets of eight or more, to form a musical scale, and were probably played in the company of string and wind instruments. The section is a flattened ellipse, and on each side of the body appear 18 blunt spikes, or bosses, arranged in three double rows of three. These often show marks of filing, and it has been suggested that they were devices whereby the bell could be tuned to the requisite pitch by removing small quantities of the metal. The oldest specimen recovered in a closed excavation is one from P'u-tu Ts'un, dating from the 9th century BC.

Secular bronzes

Vast numbers of secular bronzes were cast. These include weapons, such as the *chih* and *ko* dagger axes and the short sword; chariot and harness fittings; trigger mechanisms for bows; weights, scales, and measures; belt hooks; and mirrors. The last appear in great numbers from the



Figure 173: Chinese mirror back, bronze with lacquer, T'ang dynasty (AD 618–907). In the Museum of Fine Arts, Boston. Diameter 21 cm.
By courtesy of the Museum of Fine Arts, Boston, Marshall H. Gould fund

5th century BC onward. They are flat disks, with a central perforated boss by which they could be mounted on a stand. Their backs are covered with a maze of intricate relief designs and feature a diversified series of well-defined subjects (Figure 173). (W.Y.W.)

Iron. Iron began to take its place in the brilliant Bronze Age culture of China during the Ch'in dynasty (221–206 BC) and the Han dynasty (206 BC–AD 220). By the end of the 2nd century AD, bronze weapons had been almost completely supplanted, and iron had been generally substituted for bronze in common use in utensils and vessels of various kinds, tools, chariot fittings, and even small pieces of sculpture. These were commonly cast in sand molds, were patterned after bronze prototypes, and were typical of the Han period in style and decoration.

From the 9th century, iron increasingly took the place of bronze in China as a material for sculpture, especially in the north and under the Sung dynasty. The few extant examples from the 11th century and later show work done on a larger scale and in coarser technique than the bronzes, though the modelling is usually more naturalistic.

Several iron pagodas, dating from the 10th to the 14th century and ranging in size from miniature models to towers 100 feet or more in height, give further evidence of the dexterity of the Chinese iron caster. The pagodas imitate, in detail, both the structural and decorative effects of the more common tile-roofed brick pagodas. Iron for temple furniture has long been in use, and a large number of the braziers, censers, caldrons, and bells found today in the temples are of iron.

Iron pictures

In China in the 17th century the iron picture was developed, the craftsmen seeking to reproduce in permanent form through the medium of wrought iron the effects of the popular ink sketches of the master painters. When completed, these pictorial compositions were mounted in windows, in lanterns, or in frames as pictures. When in the latter form, a paper or silk background often bore the signature and seal of the maker, heightening the resemblance to a painting. The craft flourished in Anhwei Province and is still practiced, though with less patience and fineness than formerly. (B.Ma.)

Gold and silver. In ancient China gold and silver were rare. Gold was used as an inlay for bronzes in the Chou dynasty (1111–255 BC), and between the 6th and the 2nd centuries, gilding and silvering were common. Dress

hooks and small items of jewelry were sometimes cast in gold and silver and imitated the more usual bronze forms. Granular work—a technique that probably has an Indian origin—was used for jewelry.

Silverwork first became important during the T'ang dynasty (AD 618–907), when the Chinese had learned from the Sāsānid Persians how to chase the silver. In the beginning, they followed their teachers very closely in the forms of the bowls and larger vessels as well as in the patterns. T'ang drinking vessels, ewers, trays, and lobed oval dishes on a stem are Persian shapes transformed by Chinese taste. Among the patterns are vine and palmette scrolls of great variety, hunting scenes, and landscapes of symmetrical flowers and trees with birds and animals; all of these have parallels in Persian silver and textiles but are more delicate in their Chinese version. The techniques used by the Sāsānid silversmiths were adopted by the Chinese; for example, double sheets for a bowl and tracing of the patterns on ring-matted ground. T'ang jewelry is made of gold or gilt silver.

During the Sung dynasty (960–1279), silverwork declined in technical quality but jewelry played a more dominant role. Hair ornaments became increasingly intricate, with elaborate naturalistic flowers and various auspicious symbols.

Revival of silverwork in the Yüan and Ming periods

During the Yüan (1206–1368) and Ming (1368–1644) periods, skill in silverwork revived, and once again the smiths followed many Near Eastern styles. Drinking vessels (ewers and cups), boxes, and even large ceremonial gold vessels have been found in Ming tombs. During the excavation of the tomb of Emperor Wan-li (1572–1620), a series of gold vessels set with precious stones was found. All of the gold items are decorated with incised patterns of dragons, phoenixes, and similar subjects.

During the Ch'ing period (1644–1911/12), both silver and gold were used lavishly, and gold filigree work especially is common in the 18th century. Most of the forms and ornaments employed, however, are borrowed from lacquer and porcelain ware; and only jewelry has its own style, rich combinations of kingfisher feathers glued to the metal.

Korea. The Chinese colonists who settled Korea during the Han empire (206 BC–AD 220) first brought goldsmiths and silversmiths to Korea. By the 5th to 6th century AD Korean work, as exemplified by large gold crowns and various pieces of jewelry excavated from tombs at Kyŏngju, was beginning to develop distinctive characteristics. At the time of the Unified Silla (668–635) and Koryŏ (935–1392) kingdoms, Chinese influence was strong, but the Korean style persisted in silverwork and goldwork. Several vessels with floral patterns in relief are preserved from these periods. (B.V.Gy.)

Japan. *Iron.* The Iron Age in Japan is supposed to have begun in the 2nd century BC, though the chief early remains are weapons that date from the dolmens of the 2nd to the 8th century AD. The Japanese iron founder attained a considerable skill at an early date and acquired a social position that was never attained by the bronze caster or by the ironworkers in China, where the Bronze Age tradition was much stronger. It is apparent that iron was used in China chiefly as a substitute or imitative medium; it was worked often with great skill but with little artistic invention. In Japan, however, the ironworker developed a distinctive and original means of expression and high artistic attainment in accessories for the sword. With the rise of feudalism and the establishment of the samurai class after the wars of the 12th century, the necessary equipment of the warriors became a focus for the efforts of the artist.

At first these efforts were devoted to the embellishment of defensive armour, but from the 15th century the sword became the centre of attention. The blade is not properly part of the subject of this article; but in the mountings, especially the guards (*tsuba*), is found exquisite artistry expressed chiefly in iron. A remarkably soft and pure variety of the metal especially free from sulfur was employed. It was worked by casting, hammering, and chiselling; and innumerable surface effects were obtained by tooling, inlaying, incrustation, combination with other metals, and patination by various, usually secret, processes. Simple conventional patterns, crests, and pictorial designs were the bases for the decoration. As these were often furnished by painters or designers, the criterion of connoisseurship in Japan is the unsurpassed technical quality of the handling of the iron itself. With the promulgation of the edict of 1876, prohibiting the wearing of swords, this art came to an end, but the skill of the Japanese ironworker may still be noted in numerous small decorative objects. (B.Ma.)

Gold and silver. Knowledge of metalwork seems to have spread to Japan by way of Korea during the Yayoi period (*c.* 250 BC–*c.* AD 250), but gold and silver never played any important role there. In the Nara period (AD 710–784), the Chinese T'ang style was dominant, and most of the goldwork and silverwork preserved in the Shōsōin at Nara was made under Chinese influence or by Chinese workmen. Silver vessels were used extensively among the aristocracy in the Heian period (794–1185), though not many of these vessels have survived, and both gold and silver were often used for applied reliefs or as inlay on bronze. In the later periods the use of precious metals was practically confined to inlays in bronze or iron, and the highest technical skill is shown by the artists who made the sword fittings. (B.V.Gy.)

AMERICAN INDIAN PEOPLES

Pre-Columbian. In pre-Columbian America, gold, silver, and copper were the principal metals that were worked, with tin, lead, and platinum used less frequently. When the Spaniards arrived in the New World in the 16th century, they found a wide range of well-developed technical skills in fine metalwork in Mexico, Costa Rica, Panama, and the Andean region. They had very little to offer the Indian smiths, who had already mastered the techniques of cold hammering and annealing; embossed decoration and chasing; pressing sheet gold over or into carved molds to make a series of identical forms; sheathing wood, bone, resin, and shell ornaments with gold foil; decorating with metal inlays and incrustation with jade, rock crystal, turquoise, and other stones; joining by clinching, stapling, and soldering; possibly drawing gold wire (in Ecuador and western Mexico); casting by the lost-wax method of solid and hollow ornaments, often with false filigree or false granulation decoration; wash gilding; and colouring alloys containing gold by "pickling" in plant acids. There was some regional specialization: hammer work in "raising" a vessel from a flat disk of sheet gold or silver reached its apogee in Peru (Figure 174), and lost-wax casting was highly developed in Colombia, Panama, Costa Rica, and Mexico. Miniature, hollow lost-wax castings of the Mixtec goldsmiths in Mexico have never been surpassed in delicacy, realism, and precision; and some

Metalwork skills of pre-Columbian Indians

D.T. Easby, Jr.

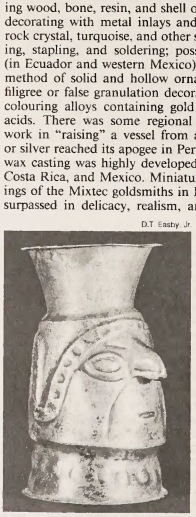


Figure 174: Peruvian silver effigy beaker, raised from a flat sheet of metal, pre-Columbian, AD 1200–1400. In a private collection, Philadelphia. Height 12.1 cm.

solid-cast frogs from Panama are so tiny and fine that they must be viewed through a magnifying glass to be appreciated. In Mexico bimetallic objects of gold and silver were made by two-stage casting; the gold part was cast first and the silver, which has a lower melting point, was then "cast on" to the gold in a separate operation. (A famous example is the pectoral of Teotitlán del Camino in the National Museum in Mexico City.) A silver llama in the American Museum of Natural History in New York City indicates that the Peruvian smiths had taken the first step toward cloisonné, the cloisons being filled with cinnabar instead of enamel.

A truly great technological and artistic triumph of the pre-Hispanic workers in Ecuador was the making of complex beads of microscopic fineness from an alloy of gold and platinum. This feat was achieved by sintering (to combine by alternately hammering and heating without melting) gold dust and small grains of alluvial platinum. (Platinum was a metal not to be used in Europe until 500 or 600 years later.)

Social standing of pre-Hispanic goldsmiths

As in other early cultures, the pre-Hispanic goldsmiths were a privileged and highly respected group, sometimes having their own patron deity such as Xipe Totec in Mexico or Chibchachun in Colombia. In Peru just before and at the time of the Conquest, the goldsmith (*kori-camayoc*) is said to have been a full-time government worker, who was supported by the state and who produced exclusively for the Inca. According to early Mexican picture writings (codices) and accounts of the Spanish chroniclers, the craft was hereditary, the secrets passed on from father to son.

The earliest examples of metalwork in the New World come from the "Old Copper" culture that flourished in the upper Great Lakes region of North America beginning about 4000 BC and continuing over the course of the next 2,000 years. The earliest goldwork is considerably later and consists of sheet-gold adornments with embossed decoration from Chongoyape, Peru, that were made sometime between 1000 and 500 BC. Casting seems to have begun in Mochica times early in the Christian Era in northern Peru, whence it is thought to have spread northward into Ecuador, Colombia, Panama, Costa Rica, and finally Mexico. Dating in the intervening areas is problematical, but it is generally accepted that fine metalwork in gold, silver, and copper did not reach the valley of Oaxaca in Mexico until about AD 900. Some finds in western Mexico suggest an earlier beginning date there and also that knowledge of the craft came by sea rather than overland from South America.

It is said that the Spaniards saw some pre-Columbian goldwork when they first arrived in Florida, but none seems to have survived. Some pre-European North American copper work, however, has survived. Metalwork was limited to a few regions in pre-European times. The "Old Copper" culture people took advantage of deposits of native copper (as opposed to smelting copper ores) to make tools and implements, and at a later period the Hopewell people extensively made copper ornaments and weapons, produced by cold hammering. A few copper bells also have been found in Arizona Hohokam sites, but these are imports that were manufactured in Mexico.

Southwest Indian. The famed Indian silverwork in the southwestern United States did not begin until 1853, when the craft was introduced to the Navajo by Mexican smiths. Although the origin is Mexican, certain ornament types and modes of decoration among the Navajo, as one scholar points out, trace back to earlier Indian silverworking in the eastern woodland, the plains, and the Rocky Mountains. It was not until 1872 that the first Zuni smith learned the craft from the Navajo. The Zuni had been carving turquoise long before the introduction of silversmithing, so it is not surprising that the most prominent characteristic of Zuni work is the extravagant use of turquoise insets. Navajo work is distinguished by die-stamped designs, whereas die work is very rare in Zuni silver. Authentic Navajo and Zuni pieces are still being made, but the tourist market has been flooded with cheap, commercial imitations.

Modern. The outstanding centre for fine handwork in silver in the Western Hemisphere is the little village of Taxco in the state of Guerrero, Mexico. An American resident, William Spratling, revived the ancient craft there in 1931 and trained a whole generation of talented silversmiths. (D.T.E.)

AFRICAN NEGRO PEOPLES

In Africa jewelry was fashioned from gold and silver as well as from nonprecious metals; heavy neck rings, anklets, and bracelets, for example, were made of forged iron or cast brass. Except for iron, metals were usually associated with prestige and/or leadership. Metals were also used for utilitarian objects such as Ashanti cast-brass weights (for weighing gold dust), which depict humans, other animals, vegetables, and geometric forms. The Nupe were excellent metalworkers, manufacturing a variety of vessels decorated with embossed designs.

Iron jewelry

Throwing knives of the Congo, often with punchwork designs, exemplify finely forged, abstract forms of iron weapons. Blacksmiths produced such ritual utensils as single or double gongs; Bambara, Dogon, and Lubi staffs topped with equestrian, human, and animal figures; and Yoruba and Benin shrine pieces containing mammal and bird forms.

Brass figure sculpture, which was cast by the lost-wax process, was usually the prerogative of royalty, as in Dahomey, and at Ife and Benin in Nigeria. Ife castings appear quite naturalistic and are among the finest sub-Saharan art. They are mostly hollow-cast heads, possibly used in ancestral rites. Benin "bronzes" were reported as early as the 16th century, but not until the 1890s did they become well-known in Europe. Local traditions indicate that the technique and the first caster came from Ife, in the 14th century. Predominant forms were heads representing deceased Benin kings (Figure 175), often supporting a carved

By courtesy of the Museum für Volkerkunde, Vienna, photograph © Photo Meyer K.G.



Figure 175: Memorial king's head, bronze, from Benin, Nigeria, 15th century. In the Museum für Volkerkunde, Vienna. Height 25.5 cm.

ivory tusk. These, with other figurative castings as well as bells, were placed on altars that were dedicated to early kings. Figurative plaques were used as architectural decoration. Excellent thinly cast pieces, fairly close to the style of Ife, gave way to heavy, overdecorated pieces of the later 19th century.

Although metals appear throughout Africa, the cast "bronzes" (often brass) of Nigeria are particularly noteworthy. The earliest, from Igbo Ukwu, may be as early as the 9th century, those of Ife as early as the 12th century; Benin castings are later, and those of the Yoruba most recent. Lower Niger Bronze Industries is a term referring to one or more as yet inadequately studied traditions of uncertain date from various places in southern Nigeria. (Ro.Si.)

ENAMELWORK

Enamel, in art, is a vitreous glaze or a combination of vitreous glazes fused on a metallic surface. The general term enamels is applied to metal objects that have this material as the principal decoration.

Enamels have been used to decorate the surface of metal objects, perhaps originally as a substitute for the more costly process of inlaying with precious or semiprecious stones but later as a decorative medium in their own right. Whereas paint on metal has a short life and, even when new, is overshadowed by the brilliance of the polished metal, enamelling gives the surface of metal a durable, coloured, decorative finish. With the painted enamels of the Renaissance and the portrait miniatures of the 17th century, the technique reached its most ambitious and artistic form; for here the craftsman attempted to create a version of an oil painting, using a metal sheet instead of a canvas and enamels instead of oil paints. This medium undoubtedly has its limitations—few painted-enamel plaques of the Renaissance, for example, are much more than one foot square—but while oil paints on canvas eventually fade and darken, the colours of enamels are permanent. Relatively few creative artists of distinction have chosen to work in this medium, however, and it has tended to be purely decorative.

Few types of metal objects have not, at some period, been enriched with enamelled decoration. Throughout history, jewelry has been made more colourful by the application of enamels. Similarly, arms and armour, horse trappings, and even domestic items, such as mirrors and hanging bowls, were embellished with enamel decoration. Throughout the Middle Ages, both secular and ecclesiastical objects, such as chalices, cups, reliquaries, caskets, crosiers (a staff carried by bishops and abbots as a symbol of office), and spoons, were elaborately enamelled. With the advent of painted enamels in the Renaissance, tableware was completely covered with enamel, and painted-enamel panels were used to decorate the ceilings and walls of rooms in the châteaus of France. Following upon the invention of the domestic table clock and of the watch in the 16th century, enamelling became one of the most popular forms of decoration for the dials and cases; by the 18th century, items of the drawing room, such as snuffboxes, etuis (cases for small articles like scissors and needles), tea caddies, candlesticks, scent bottles, and thimbles, were frequently made of enamel.

Objects typically decorated with enamelwork

Among the objects decorated with enamels in east Asia are vases, incense vessels, teapots, suits of armour, and sliding doors.

This section treats the material, techniques, and history of enamelwork.

Materials and techniques

Enamel is a comparatively soft glass, a compound of flint or sand, red lead, and soda or potash. These materials are melted together, producing an almost clear glass, with a slightly bluish or greenish tinge; this substance is known as flux or frit—or, in France, *fondant*. The degree of hardness of the flux depends on the proportions of the components in the mix. Enamels are termed hard when the temperature required to fuse them is very high; the harder the enamel is, the better it will withstand atmospheric agencies, which in soft enamels first produce a decomposition of the surface and ultimately cause the breakup of the whole enamel. Soft enamels require less heat to fire them and consequently are more convenient to use, but they do not wear so well, especially if subjected to friction.

Clear flux is the base from which coloured enamels are made, the colouring agent being a metallic oxide, which is introduced into the flux when the latter is in a molten state. The brilliance of an enamel depends on the perfect combination of its components and on maintaining an equal temperature throughout its fusion in the crucible. The colour of many enamels is achieved by a change in the proportion of the components of the flux rather than by a change in quantity of the oxide. For example, turquoise-blue enamel can be obtained from the black oxide of copper by using a comparatively high proportion of carbonate of soda; in the same way, a yellowish-green enamel can be obtained from the same black oxide by increasing the proportionate amount of red lead.

Clear flux is also used to make opaque enamels; the addition of calx, a mixture of tin and lead calcined, renders translucent enamels opaque. White enamel is produced by adding stannic and arsenious acids to the flux, the quantity of the acid affecting the density, or opacity, of the enamel.

Preparation of the enamel

The heated enamel, after being thoroughly stirred, is usually poured out onto a slab and allowed to solidify into cakes of approximately four to five inches (10 to 13 centimetres) in diameter. For use, each cake must be pulverized into a fine powder with a pestle and mortar; the powder then has to be subjected to a series of washings in distilled water until all the floury particles are removed. The metal, on which the powdered enamel is to be spread, is cleansed by immersion in acid and water. All trace of the acid is then removed by washing and by drying in warm oak sawdust. After the wet powder has been spread on the metal, it is allowed to dry in front of the furnace before it is carefully introduced into the muffle of the furnace (a compartment protected from the flame), where it is heated to the point at which it fuses and adheres to its metal base. The firing of enamel takes only a few minutes, and the object is then withdrawn and allowed to cool.

The various techniques practiced by craftsmen in the past differ mainly in the methods employed in preparing the metal to receive the powdered enamel.

Cloisonné. In the cloisonné technique, thin strips of metal are bent and curved to follow the outline of a decorative pattern; they are then attached, usually soldered, to the surface of the metal object, forming miniature walls that meet and create little cells between them. Into these cells, the powdered enamel is laid and fused. After it has cooled, the surface can be polished to remove imperfections and to add to the brilliance. The cloisonné technique is particularly suited to objects made of gold, such as jewelry.

Champlevé. This process is the opposite of the cloisonné technique: instead of building up on the surface of the metal object, the surface is gouged away, creating troughs and channels separated by thin ridges of metal that form the outline of the design. The troughs are filled



Figure 176: Millefiore glass and champlevé enamel on a hanging bowl in the Sutton Hoo burial ship. Anglo-Saxon, c. 625–660. In the British Museum. Diameter 6 cm.
By courtesy of the trustees of the British Museum

with powdered enamel and fused. The champlevé technique requires a thick metal base and therefore is used on copper and other base metals (Figure 176).

Basse-taille. This technique is a sophisticated extension of the champlevé method, for again the metal surface has to be cut away and filled with enamel, but here there are two major differences. First, within the area that has been cut away to receive the enamel, a design or figural composition is chased (chiselled), or sometimes engraved, in low relief. Because the highest point of the relief is below the general surface of the surrounding metal, the enamel, which is level on its outer surface, lies in varying thicknesses over the modelled surfaces of the low relief. Second, because the coloured enamels used in this technique are translucent, the composition of the low relief shows through; and, since the metal used is normally gold or silver, the light is reflected back through the translucent enamelling, adding a brilliant tonal quality to the enamel, just as sunlight enhances the beauty of a stained-glass window. The effect of the reflected light varies according to the thickness of the enamel lying over the undulating surfaces of the low relief; consequently, an impression of plasticity and of three-dimensional modelling is created by the subtle variations in tonal strengths of the enamel colours, which range from bright highlights to the rich tones of the deep recesses.

Stained-glass window in miniature

Plique-à-jour. The plique-à-jour technique is designed to produce an effect of a stained-glass window in miniature through the use of translucent enamels. The technique is exactly the same as cloisonné enamelling except that the strips of metal forming the cells are only temporarily attached—not soldered—to a metal base to which the enamel will not stick. After the enamel is fused and sufficiently annealed, the metal sheet, usually aluminum-bronze, is removed with a few light taps, leaving a network of metal strips filled with enamel "windows." The enamels can be carefully polished to enhance their appearance.

Encrusted enamelling (émail en ronde bosse). Encrusted enamelling is the term used to describe the technique of enamelling the irregular surfaces of objects or figures in the round or in very high relief. Both opaque and translucent enamels are applied to these small-scale sculptural objects, which are usually made of gold (Figure 177). The great technical problem is to devise methods of supporting and protecting these objects during the firing. Frequently, plaster of paris is used to envelop parts of the object, leaving exposed only those parts on which enamel is to be applied and fused.

Painted enamels. This technique differs fundamentally from the preceding five in that the various coloured enamels are not separated from each other by metal strips or ridges. Although these enamels are still applied in their wet, powdered state, the adjacent patch of coloured enamel is first allowed to dry to avoid one running into the other and so blurring the outline between them.

The metal generally used in this technique is copper. It is cut with shears into a plate of the size required and slightly domed with a burnisher or hammer, after which it is cleaned with acid and water. The enamel is laid equally over the whole surface both back and front, and then the object is fired. The first coat of enamel being fixed, the design is delineated by drawing with a needle through a layer of wet white enamel or any other that is opaque and most advantageous for subsequent coloration.

Grisaille enamels

In the case of grisaille enamels, the white is mixed with water, turpentine, spike oil of lavender, or essential oil of petroleum and painted over a dark-enamel ground. Light areas of the design are painted thickly; gray areas, thinly to allow the dark ground to tone the white pigment. The technique creates a strong contrast between light and shade, creating an impression of low relief. The scenes in grisaille are sometimes rendered more subtly by hatching, executed with a pointed tool or needle to reveal the dark enamel beneath.

In coloured painted enamels, enamel colours are spread over the grisaille treatment; when fired, parts of the surface are heightened by touches of gold, usually painted in thin lines, like hatchings. Other parts can be made more brilliant by the use of foil, over which the transparent enamels are placed and then fired.

History

ANCIENT WESTERN

The origins of the art of enamelling are uncertain. While there is archaeological evidence that glass was being made from the 3rd millennium BC in western Asia and from the 15th century BC glass vessels were undoubtedly being made in Egypt, there is no proof that enamelling on metal was practiced in either Asia Minor or Egypt until after the time of Alexander the Great (died 323 BC).

Mycenaean enamelling

Perhaps the origins of the art are to be found on Mycenaean metalwork of the 13th to 11th centuries BC. Six gold rings, excavated from a Mycenaean tomb of the 13th century BC at Kouklia (near Old Paphos), in Cyprus, are decorated with a cloisonné technique that suggests an intermediary stage between inlay and true enamelling. Scientific examination has shown that the different coloured enamels were not in the form of powder when they were inserted into the cloisons before being fired and fused together; rather they were in the form of fragments of coloured glass. Unfortunately, no report exists of any scientific examination of a more accomplished example of Mycenaean enamelling—the decoration on the gold sceptre found in a royal tomb at Kourion Kaloriziki, in Cyprus—but it is generally believed that this is true enamelling and datable to the 11th century BC.

If true enamelling existed in Mycenaean work, it would be reasonable to expect the technique to have been inherited by the Greeks and transmitted by them to the rest of Europe, perhaps by way of the colonies on the north shore of the Black Sea and in the south of Italy. Unfortunately, however, there is a long gap between the Mycenaean enamels and the Greek gold jewelry of the 6th–3rd centuries BC, which is sparingly enamelled, often having no more than touches of blue and white enamel enclosed by thin gold wire openwork (filigree).

Until recently the most ancient examples of enamelling outside Mycenaean art were said to be on ornaments discovered in a cemetery in the Kuban, close to the Caucasus, variously dated between the 9th and 7th centuries

Hirmer Fotoarchiv, München



Figure 177: "Goldenes Rössel," gold shrine with encrusted enamelling, French, c. 1403. Given to Charles VI of France by his queen, Isabella of Bavaria. In the pilgrimage church of Altötting, Germany. Height 62 cm.

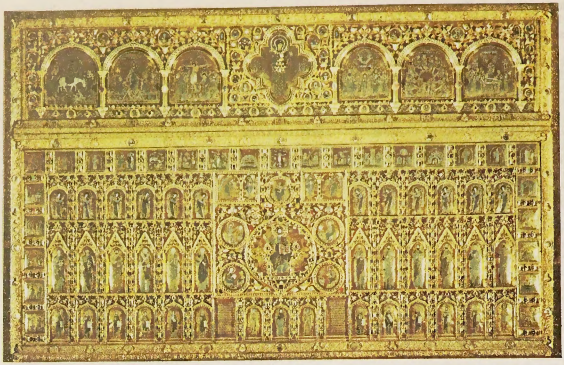


Figure 178: "Pala d'Oro," altar screen of gold cloisonné enamel, Byzantine, 10th–12th century, reassembled with later additions in a Gothic frame in 1342–45. In St. Mark's Cathedral, Venice.
SCALA—Art Resource

BC; but the most important of these Kuban enamels, the famous Maikop belt buckle (the Hermitage, Leningrad) depicting a griffin attacking a horse, is now regarded by Russian experts as a forgery. Consequently, the earliest enamelling from south Russia may date from the 3rd or 2nd century BC.

A slightly earlier date is given to a number of excavated bronze objects of western European origin, which are said to bear the remains of cloisonné enamel decoration. Until this early Celtic material has been scientifically examined and proved to be true enamel as distinct from inlaid coral, cut stone (chiefly lapis), or coloured glass applied cold, theories about it remain open to question. At the present time it is a matter of conjecture what link, if any, may have existed between the enamellers in south Russia and those Celtic craftsmen who by the 3rd century BC, if not earlier, were using red enamel in place of coral inlay.

Celtic enamel decoration

During the Roman period, enamelling—both cloisonné and champlevé on bronze—was carried on almost entirely in those old Celtic areas that had become the northern provinces of the Roman Empire. It may well be to these provincial works that a passage from the works of Philostratus, 2nd century AD, refers. The author, describing a boar hunt at which the riders appear with horse trappings ornamented in bright colours, writes:

> It is said that the barbarians in the ocean [*i.e.*, the Celtic tribes] pour these colours into bronze moulds, that the colours become as hard as stone, preserving the designs.

This is a fair description of the process of champlevé enamel and suggests that the technique, in use in the British Isles, was not practiced at the time in Greece or Italy. Enamelled horse trappings such as Philostratus describes have been found in many places in the British Isles. This type of Celtic enamelling of the Roman period lived on in northwest Europe, particularly in Ireland, until as late as the 12th century. Some of its more striking effects seem to be derived from Roman glassmaking practices, particularly its use of millefiori glass, a mosaic of very thin glass rods of different colours and shapes fused together and then cut into thin sections, which the Celtic craftsmen fused into a ground of coloured enamel.

MEDIEVAL

Byzantine. The most dramatic development in the history of enamelling took place in the Byzantine Empire between the 6th and 12th centuries, a period during which only the cloisonné technique—almost exclusively executed on gold—was in use. At their zenith in the 10th–11th centuries, Byzantine enamellers created delicate, highly expressive miniature scenes in a great range of colours that shine like jewels. The masterpiece of this period is the altar screen "Pala d'Oro" in St. Mark's, Venice (Figure 178), believed to have been brought from Constantinople to Venice about 1105. The quality of Byzantine enamelling began to decline in the late 12th century.

Islāmic. There is no direct evidence that enamelling on metal was practiced at any Islāmic centre in western Asia. Scholars who argue that the technique of Byzantine gold-cloisonné enamelling originated in Syria before the 7th century AD can point to just one object on inconclusive stylistic considerations, associated with Umayyad Syria. Only one other enamelled object has survived with strong Islāmic connections: a dish with an Arabic inscription referring to an Artuqid Prince, who reigned AD 1114–44 (Figure 179). The enamel technique is cloisonné, but with bronze wires soldered onto a copper base. As no other examples have been found and as the inscription in Arabic indicates an imperfect knowledge of the language, it may be the work of a Byzantine craftsman working in the Artuqid kingdom.



Figure 179: Copper dish with cloisonné enamelling in turquoise, cobalt blue, red, yellow, and white; the Arabic inscription refers to the Artuqid prince of Āmid and Ḥiṣn Kayfā, Dāʾūd ibn Sugmān (reigned AD 1114–44). In the Tiroler Landesmuseum Ferdinandeum, Innsbruck, Austria. Diameter 23.11 cm.
By courtesy of the Tiroler Landesmuseum Ferdinandeum, Innsbruck, Austria; photograph A. Dermenega

Western European. As early as the 7th century, according to some scholars, Byzantine work was being copied by Lombard craftsmen in northern Italy; later it was imitated in Sicily and other parts of Italy—even perhaps in England, where the famous Alfred Jewel, made to the order of the English king Alfred the Great in the 9th century AD, shows strong Byzantine influence. In the Ottonian period (AD 936–1002), gold-cloisonné enamelling seems to have flourished in eastern France, and in the Rhineland, particularly among the goldsmiths working at Essen and in the workshops of Archbishop Egbert (AD 937–993) at Trier.

In western Europe cloisonné enamelling was abandoned in the 12th century, in favour of the champlevé technique executed on a base metal such as copper or bronze. This revival may have taken place first in Spain, in the valleys of the Rhine and the Meuse, or in France at Limoges; but, by the middle of the century, expert craftsmen in these centres—and in England—had established it as one of the foremost mediums for artistic expression in the Romanesque style. In the Mosan school, the famous 12th-century enamellers Godefroid de Claire at Liège and Nicholas of Verdun created champlevé enamelwork of unprecedented merit. The best work from Limoges was executed at the turn of the 12th–13th century; thereafter, the output was commercialized and standards fell steadily throughout the 13th and 14th centuries.

Basse-taille enamelling

In the late 13th century, gold and silver objects were again decorated with enamel but in a new technique, basse-taille enamelling.The earliest surviving dated example was made in Italy in 1290. Throughout the following century, Italian goldsmiths, particularly from Siena and Florence, produced pictorial masterpieces in this medium. The technique was especially favoured in Spain and France. No more accomplished example has survived than the "Royal Gold Cup" (British Museum), commissioned by the brother of the French king Charles V about 1380. The sides and the cover have scenes depicting the life and martyrdom of St. Agnes in the most glowing rich colours and elegant draftsmanship of the period. The great era of basse-taille enamelling ended with the Renaissance, though it remained popular in Spain and southern Germany, chiefly in Augsburg, to the middle of the 17th century.



Figure 180: "The Acrobats," painted enamel plaque in grisaille with flesh tints applied as washes or brushed over in serrated thin lines of red, after a design of Juste de Juste, a sculptor working at Fontainebleau. The work was executed at Limoges, *c.* 1550. In the Walters Art Gallery, Baltimore. 43.5 cm × 30.48 cm.
By courtesy of the Walters Art Gallery, Baltimore

15TH CENTURY TO THE PRESENT: EUROPEAN

Plique-à-jour and encrusted enamelling

Under the patronage of the courts of France and Burgundy in the late 14th and first half of the 15th centuries, goldsmiths devised new and more audacious methods of enamelling. Using translucent coloured enamels, they created the effect of stained-glass windows in miniature by the technique known as plique-à-jour. One of the loveliest pieces is the silver-gilt Merode beaker of Flemish or Burgundian origin, probably *c.* 1430–40, decorated with two bands of enamels set in tiny windows with Gothic tracery (Victoria and Albert Museum, London). Employing another technique, encrusted enamelling, they created both large-scale, three-dimensional compositions and miniature work to be worn as jewelry. Among the finest and earliest surviving examples is the Reliquary of the Holy Thorn (in the Waddesdon bequest in the British Museum): the Holy Thorn, set in a gem, is surrounded by the Last Judgment scene, in which all the figures (20) are enamelled, many of them being executed wholly in the round. The taste for this type of enamelled goldsmith work spread to all the courts of Europe; and, although the style changed several times, first from Gothic to Renaissance and then to Baroque, the essential extravagant toy-like quality remained. Of all the Renaissance goldsmiths who helped to create an international style, however, only Benvenuto Cellini wrote (*c.* 1560) a technical treatise on the subject.

Although the technique of painted enamels was probably first evolved by Flemish craftsmen about 1425–50 for the Burgundian court and perhaps developed by Venetian and north Italian enamellers between 1450 and 1500, the supremacy of the Limoges workshops was established by the beginning of the 16th century. For the next 100 years, French Mannerist art found talented expression in this medium, and, enjoying court patronage, the best Limoges enamellers strove to compete with other artists in decorating the rooms of royal palaces. Painting in grisaille was finally introduced at Limoges by about 1530–40 (Figure 180).

A new dimension was given to painted enamelwork about 1620–30 by a French goldsmith, Jean I Toutin of Chateaudun, and some rival craftsmen in Blois. Their achievement was to invent a highly skillful method for fine miniature painting in enamel colours on a white-enamel ground. Since the technique was admirably suited to the current enthusiasm for portrait miniatures, artists of distinction, such as Jean Petitot, were employed by Charles I of England and the French kings to work in this medium.

Miniature painting in enamel colours

With equal artistic skill, other French enamellers decorated items of jewelry, especially watchcases; and, by the second half of the 17th century, this craft had become centred on Geneva, where it continued to flourish into the 19th century. In England, particularly in the Midlands, the Continental style of painted enamelled "toys" was copied and produced on a large scale, but the technique of transfer printing on enamel was invented in England and brought to perfection at the Battersea (London) factory during 1753–56. The design was applied to the white-enamel ground by transferring to paper, and then to the surface to be decorated, an impression from an engraved metal plate that had been brushed with enamel colours. Throughout the 17th and 18th centuries, enamellers used the technique of fine miniature painting in enamels in Germany, Holland, England, and Russia in order to produce the "toys" of the fashionable world of society.

The technique called *en résille sur verre* flourished for only about 40 years (*c.* 1600–40), and few examples have survived. Yet it required an exceptional degree of skill. The technique consists of cutting the design in a medallion of glass, usually coloured, lining the incisions with gold and filling them with variously coloured enamels. The exponents of this kind of enamelling were mainly French.

Although surviving examples are rare, there is a dis-

tinctive group of brass objects, mainly candlesticks and andirons, which have green, blue, or white opaque enamelling. These objects were made in 17th-century England (perhaps in Sussex).

Most of the early enamelling techniques have continued to be used by goldsmiths in modern times—from the Parisian makers of gold snuffboxes in the 18th century to Carl Fabergé at the beginning of the 20th. Art Nouveau jewellers, such as René Lalique, and modern artists, such as Georges Braque, Georges Rouault, and Gerda Flockinger, have kept alive the craft of enamelling and added to the multiplicity of its ingenious effects. (H.Ta.)

CHINA

Enamels do not appear to have reached China until long after they were found throughout Europe. All authorities are agreed as to the Western origin of the art, which in all probability was introduced into China by traders or by travelling craftsmen. Although by the 5th century AD the Chinese were informed as to the production of glass—an essential material for the making of enamels—and were already highly skilled in the working of bronzes and other metals, there is no evidence that the art of enamelling was practiced before the T'ang dynasty (618–907). There is in the Shōsō-in (principal storehouse) at Nara, in Japan, a silver mirror, the back of which is decorated in cloisonné (Figure 181). It is generally agreed that the mirror is of Chinese origin, dating from the T'ang dynasty, as is certainly the case with many other objects in the collection. At present, this is the only known Chinese enamelled ware made before the 14th century; but it can be safely assumed from this piece that the art of cloisonné was developed to a respectable height in the T'ang dynasty. It appears that cloisonné work was well established in China at the end of the 14th century and that Byzantine work of similar character was also so well known as to invite comparison with the native product. The former may well have served as an example for Chinese craftsmen. As one scholar points out:

Earliest known enamelwork

> The workmanship presents occasionally . . . striking resemblances with certain enamels of the Byzantine school; the mixture of different enamels inside the wall of the same cell, the employment of gold incrustations in the treatment of the fingers and the hands, etc.

Active trade and cultural intercourse between the Near East and China during the Yüan (Mongol) dynasty must have been the reason for this revival of enamelwork, which then flourished through the Ming and Ch'ing dynasties (1368–1644 and 1644–1911/12, respectively).

Chinese enamels fall into three categories—cloisonné, champlevé, and painted. In none does the technique vary appreciably from that employed in Western countries.

Cloisonné. The earliest example of cloisonné enamel that can be authentically associated with east Asia is the silver mirror in the Shōsō-in mentioned above. Its cloisonné back is decorated in a design of a six-petalled blossom in three layers, the tips of the outer rows of petals forming 12 points of the mirror. The piece is regarded as a T'ang dynasty work. Apart from this, the sequence of known Chinese enamels begins in the Yüan period, and the earliest recorded marks belong to the reign of the last emperor of that dynasty (1333–68). The great period of production is certainly that of the Ming dynasty, which followed.

Ming enamels

The mark most commonly found within this period is that of the Ching-t'ai reign (1449–57). The Ming enamels, bold in design with fine depth and purity of colour, were never surpassed in later epochs. The two shades of blue, a dark-lapis-lazuli tone and a pale sky blue with a very slight tinge of green, are particularly excellent. The red is of dark-coral tint, and the yellow full-bodied and pure. Greens derived from copper are sparingly used. Black and white are the least successful, the former shallow and dull, the latter clouded and muddy. As fine as the Ming enamels are, however, there is an imperfection of technique, close examination revealing minute pitting in the enamels, which was caused by inadequate packing of the material, and some lack of polish in the surface. These technical defects, however, do not appreciably detract from the great artistic value of the Ming enamels (Figure 182).

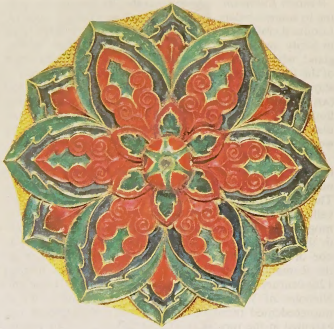


Figure 181: Mirror back, cloisonné enamel on silver, Chinese, T'ang dynasty, 9th century. In the Shōsō-in, Nara, Japan. Diameter 18.7 cm.
By courtesy of the Shosoin, Nara, Japan

A great revival of art industries took place under the patronage of the emperor K'ang-hsi (1661–1722), who, in 1680, established a series of imperial factories. He commissioned sets of incense vessels of cloisonné enamel for presentation to the numerous Buddhist temples founded under his auspices in the neighbourhood of Peking, as well as other objects for the honorific gifts that were characteristic of his enlightened reign. The enamels produced during his time are marked by an improvement in technical quality as compared with those of the Ming period; to



Figure 182: Ming dynasty vase, cloisonné enamel, Chinese, *c.* 1500. In the British Museum. Height 41.5 cm.
By courtesy of the trustees of the British Museum

a considerable extent they also retained the finer qualities of the Ming wares. In many cases the forms of ancient bronze vessels were revived and enriched with enamels.

The style of this reign persisted during that of K'anghsi's successor, Yung-cheng (1722–35), while the long rule of Ch'ien-lung (1735–96) was marked, in enamel as in the case of many other industrial arts, by a further perfection of technique but by a loss of much of the vigour of design and breadth of execution that distinguished the products of earlier periods. Modern enamels, although they are primarily imitations of older work, are more hurriedly made and therefore not so well finished as the older work.

Champlevé. Some of the most ancient enamel examples extant belong to this class, and examples employing both champlevé and cloisonné are not uncommon.

Although the opaque enamels were more common, Chinese artisans occasionally used translucent enamels on a silver or gold base. The cloisonné back of the silver mirror in the Shōsō-in, for example, is decorated with transparent enamels, but important pieces such as this are rare. This technique more often appears in Chinese jewelry.

Canton enamels

Painted enamels. The painted enamels of China, generally known, from the principal seat of their manufacture, as Canton enamels, are practically identical in technique with the Limoges and other painted enamels of Europe. Specimens of the latter are known to have been taken to China by the missionaries of the late 17th and 18th centuries; they not only exercised direct influence on the Chinese ware but also, in some cases, were copied. Representations of European subjects, copies of engravings and armorial decorations, are also found. Painted enamels are termed by the Chinese *yang-tz'u* ("foreign porcelain"), the palette of colours used being the same as with enamelled porcelain, whose decoration under foreign influence is called *yang-ts'ai* ("foreign colours"). A ground of opaque enamel, generally white, is laid on the copper, and on this the colours are superimposed and fired. The best period of this art was the 18th century. Although imitations have continued to be made, nothing of real quality in this style was produced after the termination of the reign of Ch'ien-lung in 1796. The method has always been looked upon by the Chinese as alien in taste; in fact, a great part of the Canton enamels were made for export, not only on European commission but also for clients in India, Persia, and several other Asian countries.

JAPAN

The art of enamelling in Japan may date back to the 7th century. One enamelled metal piece, discovered in a tomb near Nara, probably from the 7th century, seems to be of Japanese origin. The civic Taihō code, compiled in the 8th century, provides one official to be in charge of founding metals and "painted glass decoration." Subsequently, however, this art seems almost to die out until the 17th century. When Dōnin Hirata I (1591–1646) made enamelled wares, having learned the technique from Koreans, his art was highly appreciated by Tokugawa Ieyasu, then the shogun of Japan, under whose patronage Hirata worked in Kyōto. There is a suit of armour with enamelled metal fittings ascribed to him, as well as enamelled metal fittings decorating sliding doors and lintels in the Katsura Palace, Kyōto. His family continued the trade until the late 19th century, making use, on a small scale, of both the cloisonné and champlevé methods. There was no further development of importance until Kaji Tsunekichi (1803–83) and his pupils established in Nagoya a successful manufacture of cloisonné, which obtained a considerable vogue, especially among foreigners.



Figure 183: Panel of Mt. Fuji seen through clouds, done in lineless cloisonné by Namikawa Sōsuke, 1893. In the Tokyo National Museum. 63 cm × 113.6 cm.
By courtesy of the Tokyo National Museum

Lineless cloisonné

While Kaji used brass cloisons and opaque enamel colours only, his successors used silver cloisons and succeeded in making both transparent and translucent enamels. They further modified the cloisonné process with remarkable ingenuity and produced work of great interest. First, they reproduced in minutely detailed cloisonné work realistic pictures, of trees and flowers, for example. This effort led them in the 1880s to produce lineless cloisonné enamels, which have all the beauty and brilliance of true cloisonné, with thick layers of enamel colours, but which, showing no trace of cloisons, permit a gradation of colours as in the less clear and brilliant painted enamels. The effect was achieved by taking off the cloisons before each firing, the process being repeated at least three times. The artists working for the factory of Namikawa Sōsuke of Tokyo in the late 19th century were most successful in this technique (Figure 183). Another Namikawa of Kyōto worked in true cloisonné. The factory of Jubei Ando of Nagoya has produced more variations. These developments have carried the art of enamel very far from the old traditions; while the skill and ingenuity of technique they evince may be appreciated, there is a danger of losing the artistry peculiar to the art of enamelling. (Ed.)

LACQUERWORK

Lacquerwork consists of certain metallic and wood objects to which coloured and frequently opaque varnishes called lacquer are applied. The word lacquer is derived from lac, which is the basis of some lacquers. The lacquer of East Asia, China, Japan, and Korea must not be confused with other substances to which the term is generally applied; for instance, the lac of Burma, which is the gummy deposit of an insect, *Coccus lacca,* and the various solutions of gums or resin in turpentine of which European imitations of Eastern lacquer have been and are concocted.

This section deals with the materials, techniques, and history of artistic lacquer ware.

Techniques

OBTAINING AND PREPARING LACQUER

Lacquer, as used in China and Japan, is a natural product, the sap of a tree, *Rhus vernicifera;* subject to the removal of impurities and excess water, it can be used in its natural state, though it was frequently adulterated. The tree, which is indigenous to China and has certainly been cultivated in Japan at least since the 6th century AD, is tapped at about the age of 10 years, when lateral incisions are made in the bark and the running sap is collected during the months of June to September. Branches of a diameter of one inch (about three centimetres) or more are also tapped, the bark having first been removed. Smaller branches are cut off and soaked in water for 10 days, and the sap is collected, producing a lacquer (*seshime*) of particular quality, used for special purposes. These processes kill the tree, but the wood, when of sufficient size, is of some use for carpentry. From the roots five or six shoots spring up, which become available for the production of lacquer after about six years, and the operation can be thus continued for a considerable length of time before the growth is exhausted. The Chinese and Japanese methods

Collecting the sap

are practically identical in this respect, but the cultivation of the tree does not seem to have been as systematic in China as in Japan.

The sap is white or grayish in colour and about the consistency of molasses. On exposure to the air it turns yellow-brown and then black. It is strained through hempen cloth to remove physical impurities, after being pounded and stirred in shallow wooden tubs to give it uniform liquidity. It is then slightly heated over a slow fire or in hot sunshine, and stirred again to evaporate excess moisture, and stored in airtight vessels.

The basis of lacquer ware, both in Japan and in China, is almost always wood, although it was also occasionally applied to porcelain, brass, and white metal alloys. In some instances, objects were carved out of solid lacquer. The wood used, generally a sort of pine having a soft and even grain, was worked to an astonishing thinness. The processes that follow are the result of extraordinary qualities of lacquer itself, which, on exposure to air, takes on an extreme but not brittle hardness and is capable of receiving a brilliant polish of such a nature as to rival even the surface of highly glazed porcelain. Moreover, it has the peculiar characteristic of attaining its maximum hardness in the presence of moisture. To secure this result, the Japanese place the object in a damp box or chamber after each application of lacquer to the basic material (wood, etc.). The Chinese are said (in an account of the industry dating from 1621–28) to use a cave in the ground for this purpose and to place the objects therein at night in order to take advantage of the cool night air. It may, indeed, be said that lacquer dries in a moist atmosphere.

Hardening with moisture

APPLICATION

The joiner's work completed and all knots or projections most carefully smoothed away, cracks and joints are sealed with a mixture of rice paste and *seshime* lacquer, until an absolutely even surface is obtained. It is then given a thin coat of *seshime* lacquer to fill up the pores of the wood and to provide a basis for succeeding operations, which may number as many as 20 or 30 or more, of which any one of the following may be taken as typical. On the basis, as above described, is laid a coat of lacquer composition which is allowed to harden and is then ground smooth with whetstone. Next comes a further coat of finer composition, in which is mixed some burnt clay, which is again ground and laid aside to harden for at least 12 hours. On this is fixed a coat of hempen cloth (or, rarely in Japan but more often in China, paper) by means of an adhesive paste of wheat or rice flour and lacquer, which needs 24 hours at least to dry. The cloth is smoothed with a knife and then receives several successive coats of lacquer composition, each demanding the delay necessary for hardening. On this is laid very hard lacquer, requiring a much longer drying interval, afterward being ground to a fine surface. Succeeding coats of lacquer of varying quality are now laid on, dried, and polished.

This preliminary work, requiring for artistic lacquer at least 18 days, produces the surface on which the artist begins his task of decoration. A large number of processes have been at his command, especially in Japan, but the design was first generally made on paper with lacquer and transferred while still wet or drawn directly with a thin paste of white lead or colour. In carrying it out the artist used gold or silver dust applied through a quill, a bamboo tube, or, for equal distribution, a sieve. Larger fragments of the precious metals were applied separately by hand, with the aid of a small, pointed tool. In one typical instance approximately 500 squares of thin gold foil were thus inserted within one square inch (six square centimetres). These processes each entailed prolonged hardening periods and meticulous polishing. Relief was obtained by modelling with a putty consisting of a mixture of lacquer with fine charcoal, white lead, lampblack, etc., camphor being added to make it work easily. Lacquer was sometimes engraved, both in China and Japan.

CHINESE CARVED LACQUER

The carved lacquer of China (*tiao-ch'i*), which was imitated but never equalled in Japan (as the Chinese have never reached the perfection of the Japanese gold lacquer ware), needs particular notice (Figure 184.) In this the lacquer was built up in the method described above, but to a considerable thickness; when several colours were used, successive layers of each colour of uniform thickness were

Use of successive layers of different coloured lacquer

By courtesy of the Victoria and Albert Museum, London; photograph, A.C. Cooper Ltd



Figure 184: Imperial Chinese throne of the Ch'ien-lung emperor (1736–96), red lacquer carved in dragons and floral scrolls, Ch'ing dynasty. In the Victoria and Albert Museum, London. 1.19 m × 1.26 m × 91 cm.

arranged in the order in which they were to predominate. When the whole mass was complete and homogeneous, it was cut back from the surface to expose each colour as required by the design. When the lacquer was cold and hard, the carving was done with a V-shaped tool kept very sharp. The cutting was done with amazing precision—no correction of faults was possible, for each layer had to be exactly and accurately reached and the final result precisely foreseen from the beginning of the work. The red lacquer (*t'i-hung*), so well known and justly appreciated, was coloured with cinnabar (red mercuric sulfide). Other colours include a deep and a lighter olive-green, buff, brown, black, and purple (aubergine).

JAPANESE PROCESSES

In Japanese lacquer, the following are the chief processes used: *nashiji* (pear skin), small flakes of gold or silver sunk to various depths in the lacquer; *fundame,* fine gold or silver powder worked to a flat, dull surface; *hirame,* small, irregularly shaped pieces of sheet gold or silver placed on the surface; *togidashi,* the design built up to the surface in gold, silver, and colours with many coats of lacquer and then polished down to show them (Figure 185); *taka-makie,* decoration in bold relief; *hiramaki-e,* decoration in low relief; *rō-iro,* polished black; *chinkin-bori,* engraved lacquer; *kirikane,* square dice of sheet gold or silver, inserted separately on the surface; and *raden,* inlaid shell and metal. From the earliest recorded times shell was used in the adornment of lacquer in China as well as in Japan, being inlaid on the surface in patterns as well as in small squares like *kirikane* and dust. For this purpose various shells were used, mother-of-pearl (for larger work), nautilus, pear shell, sea-ear (*Haliotes,* Japanese *Awabi*) and *Turbo cornutus* (Japanese *Sazae*). For a very charming form, called *laque burgauté,* the iridescent blue and green shell of the sea-ear was delicately engraved with gold and silver as early as the Ming period (1368–1644) and also in Japan. Chinese lacquer was also inlaid with jade, malachite, coral, soapstone, ivory, porcelain, and other substances.

Laque burgauté

Historical development

CHINA

The use of lacquer in China goes back traditionally to legendary times. A late Ming manuscript, the *Hsiu shih*



Figure 185: Japanese box with "ardent lover" theme in *togidashi* on a *rō-iro* background, signed Katsukawa Shunshō, early 19th century. A young woman takes black tooth stain from a bowl held by her attendant and squirts from her lips the characters for "perseverance in love." In the Victoria and Albert Museum, London. 21.6 × 25 × 5 cm.
By courtesy of the Victoria and Albert Museum, London: photograph, A.C. Cooper Ltd

lu, states that it was first employed for writing on bamboo slips, then for utensils for food, made of black lacquer, and subsequently for vessels for ceremonial use, of black with red interiors (Figure 186). During the Chou dynasty (1111–255 BC) it served for the decoration of carriages, harnesses, bows and arrows, etc., and was the subject of official regulations. At this time, gold and colours are said to have come into use. About the 2nd century BC buildings were decorated with lacquer, as were musical instruments. Under the Han dynasty (206 BC–AD 220) further development took place. Pot covers of paper covered with lacquer, found near Port Arthur, are attributed to this period.

Of the lacquer of the T'ang dynasty (618–907) more reliable information is available, for the collections preserved in the Hōryū-ji in Japan, founded AD 607, and those assembled by the Japanese emperor Shōmu (724–748), deposited after his death in the Imperial treasury (Shōsō-in) at Nara, contain many objects of Chinese origin; in particular, musical instruments with inlay of cutout figures of gold and silver inserted on the surface, covered with lacquer, which was then rubbed down until the metal ornaments were again visible.

Under the Sung dynasty (960–1279) the industry further developed, and the use of gold and silver lacquer in the utensils made for the palace is particularly recorded. The chief centres of manufacture were Chia-hsing and Su-chou. A lacquer box of the early Sung period, probably once of rhinoceros-horn colour, black and red, with gold dust and silver wire, is one of the very few known examples of the period. Toward the close of the period (*c.* 1220) it is stated that lacquer wares were exported from Ch'uan-chou, Fukien to Java, India, Persia, Japan, Mecca, and other places. Chinese writers record the existence of carved red lacquer during the time of the Yüan dynasty (1206–1368) as well as of pierced ware and that inlaid with shell.



Figure 186: Wood bowl decorated in red and black lacquer with stylized birds and animals, from Ch ang Sha, China, late Chou dynasty, 3rd century BC. In the Seattle Art Museum, Washington. Diameter 25 cm.
By courtesy of the Seattle Art Museum, Washington

Ming lacquerwork

Of the state of the industry under the Ming dynasty (1368–1644) there are contemporary Chinese descriptions; for instance, in the *Kō ku yao lun*, published in 1388, the *Ch'ing pi-ts'ang*, published in 1595, and the *Hsiu shih lu*, which has been handed down in manuscript. This last work was written by a celebrated lacquerer, Huang Ch'eng, and bears a preface by Yang Ming, another lacquerer, dated 1625. The work itself was probably written towards the end of the 16th century. From these works one can ascertain the excellence of the carved lacquer made during the reigns of the Yung-lo (1402–24) and Hsüan-te (1425–35) emperors. Examples of carved lacquer that can be attributed to both these reigns are extant. They are bold in design and free from the superabundance of small detail that characterized later productions; the colour also is generally deeper and richer than that of the 18th-century pieces. In the 16th century there were special factories for carved lacquer at Ta-Li in Yünnan, which also produced spurious imitations. Lacquer with designs painted in gold outlines were made, early in the Ming dynasty, at Nanking and afterward at Peking, and lacquer inlaid with mother-of-pearl was made at Chi-chou in Kiangsi. In the reign of Hsüan-te, lacquer decorated in sprinkled gold was introduced from Japan and excellent copies were made by Chinese lacquerers. Toward the end of the Ming dynasty there was a decline in lacquer manufacture as a result of the troubles accompanying the fall of the last Ming emperor.

The first and perhaps the greatest of the Manchu emperors, K'ang-hsi (1661–1722), revived the lacquerwork industry in 1680, when he established a series of 27 workshops for artistic handicrafts in the precincts of the palace at Peking. Carved lacquer was, however, also made at Canton, Su-chou, and Foo-chow; and the Jesuit Louis le Comte, who arrived in China in 1687, gave a good account of the flourishing state of the industry at that time. In this connection it is worth noting that the period of K'ang-hsi was that which saw the first considerable importation of lacquer ware (and other objects of industrial art) into Europe. This led to the development of imitation lacquer applied to furniture and other objects, which were conspicuous features of the chinoiserie craze of the late 17th and 18th centuries. A screen, *c.* 1700 in the collections of Earl Spencer and R. Freemer Smith, Esq., for example, was made by command of K'ang-hsi for presentation to the Holy Roman emperor Leopold I, whose badge, the double-headed eagle, is incorporated in the design. Carved lacquer of this period, though far from negligible, hardly attains to the rich colour, breadth, and simplicity of that of the Ming period.

Importation into Europe

In technique the K'ang-hsi ware shows an advance and is generally free from the small cracks too often found to have developed in the Ming products. The perfection of this quality, apart from other considerations, is found in the lacquer ware of Emperor Ch'ien-lung (1735–96), a devoted admirer of this art, who employed it on a large scale for the furniture and fittings of his palaces (Figure 184), as well as for ceremonial and commemorative gifts. The workmanship of objects made under his auspices is brilliant in the extreme, but the colour is hard as compared to earlier work, and the design tends to a somewhat stereotyped formalism.

Still, the 18th century can hardly be called a period of decadence in the decorative arts of China: the superb execution of its productions, a characteristic that commands admiration, redeems it from adverse criticism. The downward course began in the 19th century, with loss of originality and a falling off, due to adulteration, in the quality of the material. What was left of the Imperial fac-

tories was burnt in 1869, and, though carved red lacquer was made after that date, the industry had already ceased to have artistic importance.

JAPAN

Although the earliest reference to the manufacture of lacquer accepted by all Japanese authorities is a code of law (known as Taihō code) dated 701, there can be no doubt that the manufacture was brought to Japan from China via Korea at the time of the introduction of Buddhism in the middle of the 6th century. At the same time, according to tradition, the Chinese lacquer tree was introduced. The earliest piece of lacquer known today that is accepted as having been made in Japan is the Tamamushi Shrine in the Hōryū-ji, which is attributed to the 7th century. This piece shows strong Korean influence. Many fine pieces of the late 7th and the 8th centuries, inlaid in gold, silver, or mother-of-pearl, of Chinese origin, have been preserved in the Shōsō-in, as already mentioned in the section above on Chinese lacquer. But there is one piece in the Shōsō-in, a sword-scabbard of black lacquer decorated in gold, formerly belonging to the emperor Shōmu (724–748), that is undoubtedly Japanese. This is listed in an old catalog dated 756. There are also two arrows in the Tokyo National Museum that belong to the same period. These can be regarded as the real beginnings of a Japanese style in lacquer.

The emperor Kammu (781–806) removed the capital from Nara to a new city, Heian-kyō—the modern Kyōto; and an increased luxury in the style of living brought about further developments in the art, especially in the use of gold lacquer, largely because of the spread of Buddhistic influence. This period, however, saw the beginnings of a Japanese national style as distinct from the Chinese methods and manner, imported by Buddhist missionaries. Lacquer was used at this time in the decoration of important buildings, and inlay of shell also became popular. The organization of the industry was extended, and, as early as 905, sumptuary edicts began to be issued regulating the dimensions and quantities of material to be used in the domestic utensils—chiefly of black or red polished lacquer—which now began to come into general use. From this time, it is no exaggeration to say that, to a considerable extent, lacquer filled the place occupied in China by ceramic wares. A remarkable development of this period that must not be overlooked was the production of statuary of considerable merit, made with lacquer composition (*kanshitsu*), a process derived from China but carried to a high standard in Japan for a brief period, until it was superseded by wood sculpture (Figure 187). Some few authentic examples remain of the fine lacquer of the Heian period, notably a case for Buddhist scriptures in the Ninna-ji at Kyōto, made at the beginning of the 10th century, which bears an inscription dated 919. The case is in black lacquer, sprinkled with gold dust and with a pattern of flowers, clouds, birds, and Buddhist winged genii in gold and silver *togidashi*.

Production of statuary



Figure 187: "Priest Ganjin," hollow dry lacquer (*kanshitsu*) statue, Late Nara period (724–794 AD). In the Tōshōdai-ji, Nara, Japan. Height 80 cm.
By courtesy of the Toshodai-ji, Nara, Japan

During the Kamakura period (1192–1333), in spite of the disturbance caused by the famous struggle between the Minamoto and Taira clans and the establishment of the feudal shogunate at Kamakura, which gives its name to the period, the art of making fine lacquer continued to progress under the patronage of the Fujiwara family, who maintained the Imperial court at Kyōto with ever increasing luxury. Marked features of this time are improved methods of inlay of precious metals and shell and, especially, an attractive form of design in which beautifully written poems are interwoven with the pattern (*ashide*). The process called Kamakura-*bori*, carved wood thickly lacquered with red or black, also dates from this period and continued to flourish for another two centuries or so. During this epoch occurred the beginnings of the characteristic Japanese treatment of landscape and flower subjects in design, generally in flat gold lacquer with *nashiji* and pewter inlay.

Muromachi lacquerwork

The Muromachi period (1338–1573) saw a further technical and artistic development, largely under the patronage of the shogun Ashikaga Yoshimasa (reigned 1443–73). He gave great impetus to the tea and incense ceremonies, the latter of which brought about a whole series of new applications of the art because of the exquisitely wrought small utensils required by that ritual. The ostentatious simplicity of the Zen sect of Buddhists was displayed in the use of black lacquer of the first quality with little or no ornament. Excellent work in shell inlay was also a characteristic of the time. The gold lacquer of the Muromachi craftsmen gained so great a reputation in China that artisans from that country went to Japan to learn the methods by which it was produced, though they seem to have had little success in introducing it into their own country. Among the leading Japanese craftsmen of the period may be mentioned Kōami Dōchō, Taiami, Seiami, and Igarashi Shinsai, but attribution of specific works to them is largely a matter of conjecture.

The civil wars which continuously infested Japan during the later Middle Ages checked the growth of the industry for a while, but the short Azuchi-Momoyama period (1574–1600) that followed saw at least the work of one of the greatest of Japanese artists in lacquer, Honami Kōetsu. He was the founder of a striking and original style of ornament, essentially national in character. His designs were bold and simple in detail, generally executed in high relief with masses of shell or metal inlay. The great feudal lord Toyotomi Hideyoshi (died 1598), who secured the peace of the country with a strong hand, was an enthusiastic patron of the arts, and under his patronage a real revival took place. When he died, his widow erected the Kōdai-ji at Kyōto, in which distinctive lacquer decoration called *tata maki-e (Koda-ji maki-e)* was used. This temple still contains examples of this ware that were presented by her.

The Edo school

In 1603 began the rule of the Tokugawa shogunate, which continued without a break until the restoration of the Imperial family to actual power in 1867. The first of the line, Ieyasu, established at Edo (the modern Tokyo) the great school of lacquer artists that is responsible for almost the whole of the artistic ware known outside Japan. Technical processes were still further developed with additions such as engraved lacquer (*chinkinbori*) derived from China, carved red and black lacquer from the same source, and the so-called *somada* ware of shell inlay of black, different in character from the Chinese *laque burgauté* already mentioned above.

This period also saw the introduction of the now well-known *inrō*, or portable medicine case, worn on the girdle and an indispensable addition to the national costume so long as the latter was uncontaminated by Western influence. An *inrō* consisted, as a rule, of from two to five



Figure 188: Tokugawa period writing box of black lacquered wood decorated in gold paint, inlaid lead, and pewter. Attributed to Ogata Kōrin (c. 1658–1716). In the Seattle Art Museum, Washington. 23 × 22 × 8 cm.
By courtesy of the Seattle Art Museum, Washington

compartments, beautifully fitted into each other and held together by silken cords running along each side, secured by a bead (*ojime*) and kept in place on the sash by a kind of toggle (*netsuke*), sometimes of lacquer but more often of cunningly carved wood, ivory, bone, or other material. On this class of work was lavished some of the finest artistry of the Japanese craftsmen, and the convenient size and intrinsic charm of these dainty utensils (originally, perhaps, made for seals) have caused them to be much favoured by collectors.

Chinese influence on Tokugawa lacquerware

The earlier years of the Tokugawa period saw a considerable Chinese influence in the design of lacquer, especially in *inrō;* but the work of the greatest Japanese lacquer artists, Ogata Kōrin (Figure 188), followed and extended in the late 17th century the style originated by his master, Kōetsu. Ritsuō and Hanzan in the 18th century maintained this tradition, and a considerable revival of the style took place in the early years of the 19th century, when memorial volumes of the designs of the great master were published. To the latter period belong not a few objects which have been accepted as the original work of Kōrin himself. The more formal school of lacquerers included Kōami Chōgen (1572–1607) and Komo Kitō-ye, who was appointed court lacquer artist to the shogun Iemitsu in 1636 and died in 1674. One of the most important lacquerers of the Kōami family was Kōami Nagashige (Figure 189), whose masterpiece, consisting of three connected cabinets with numerous writing cases, paper cases, and toilet cases, a mirror stand, and other accessories, was completed in 1639. It was made for the dowry of Tokugawa Iemitsu's eldest daughter, on her marriage to Mitsumoto, prince of Bitchū, whose coat of arms is affixed to all the pieces. The lacquer is now in the Tokugawa Art Museum, Nagoya. As did other craftsmen in Japan, lacquer artists followed the practice of transmitting their names to sons or selected pupils. Thus, there were ten generations of the family of Yamamoto Shunshō, who died in 1682, aged 63. The Kajikawa family continued the tradition of its founder well into the 19th century, and the same must be said of Shiomi Masanari in the 18th century, whose work was notable for the quality of the rubbed-down gold and colour lacquer called *togidashi*.

By courtesy of the Tokugawa Art Museum, Nagoya



Figure 189: Lacquered cabinet by Kōami Nagashige, 1639. In the Tokugawa Art Museum, Nagoya. 101 × 77.5 × 40 cm.

The Genroku period (1688–1703) saw, perhaps, the ultimate perfection of style and technique; but the work of the later 18th and, to some extent, of the early 19th centuries has many exquisite qualities. The later periods were characterized by more elaborate detail, but adulteration of the gold with bronze and other metallic powders was often prevalent. A fiery brown tint of the *nashiji* is a certain mark of quite late date. Nevertheless, there is plenty of good work of the 19th century, and to this period belongs the last of the great artists of the industry, Shibata Zeshin, whose work bears comparison even with some of the greatest of his predecessors, both in technique and in design.

Modern industrial conditions, however, have practically killed this ancient and beautiful art. It would not have survived so long had not the country been closed to alien influences for two and a half centuries. (E.F.S.)

EUROPE

Although East Asian objects of art were taken to Europe in considerable quantity during the 16th century, it was not until after 1600 that a real trade with China grew up, fostered by the East India companies of the Netherlands, England, and France. Porcelain and lacquerwork then became so fashionable that European craftsmen undertook to make their own. The secret of porcelain manufacture was discovered shortly after 1700, and by that time the imitation of lacquer had become established in various parts of Europe. Jesuit missionaries had cooperated with scientists and master craftsmen to create formulas, and a small body of writing had appeared, which was to be augmented in the 18th century—the golden age of European lacquer.

Origins of the European lacquerwork industry

Among the earliest surviving examples of this art is the ballot box of the Saddlers Company (Figure 190). Information on the lacquer process seems first to have been published by the Italian Jesuit Martin Martinius (*Novus Atlas Sinensis,* 1655). John Stalker and George Parker's *Treatise of Japanning and Varnishing* (London, 1688) was the first text with pattern illustrations. The English term japanning was inspired by the superiority of Japanese lacquer, which Stalker found " . . . in fineness of Black, and neatness of draught . . . more beautiful, more rich, or Majestick" than the lacquer of other places, which came to be known as "Indian" or "Bantam" work.

In France, on the other hand, *ouvrage à la Chine* was the term for the imitation of lacquer practiced at the Gobelins factory in Paris from 1672. By the end of the century Berlin had become another centre of experimentation, from which a Fleming, Jacques Dagly, brought secrets that were to lead to the 18th-century innovations of the Martin brothers: Guillaume, Simon, Étienne, Julien, and Robert. They created the lustrous *vernis Martin,* which was praised by Voltaire. The Martins decorated rooms at Versailles, and Robert's son Jean Alexandre worked for Frederick the Great II at Potsdam. French lacquer was further improved through new information provided by the *Mémoire sur le vernis de la Chine,* which the French missionary Pierre d'Incarville wrote in 1760 and which appeared as an appendix to *L'Art du peintre, doreur, vernisseur* of Jean-Félix Watin (1772), the most precise account of lacquerwork that appeared in the 18th century. In this book Watin examined the recipes of his predecessors and recommended the best formulas for lacquering objects to be used indoors, such as furniture, and outdoors, such as carriages. Although nothing could equal the excellence of Oriental



Figure 190: Ballot box of the Saddlers Company, 1619. In the collection of the Saddlers Company, London. 45 × 44 × 33 cm.

By courtesy of the Saddlers Company, London

resins, he determined that sandarac from Western juniper trees was the best substitute. This, together with various gums dissolved in alcohol and turpentine and mixed with bitumens, produced the different varnishes Watin relied upon. His book gives detailed instruction for preparing the wood, covering it with cloth, varnishing this with from eight to 20 coats, polishing the surface, drawing and painting the designs, and making relief decorations. The best results of this process (*e.g.*, an 18th-century commode attributed to René Dubois; Figure 191) were never as hard and brilliant as real Oriental lacquer, but they provided an admirable substitute; on occasions it is not easy to distinguish them, especially when East Asian designs were imitated.

Introduction of European motifs

The chief interest of European lacquer of the 18th century lies, however, in the fact that it was not a purely imitative art, as it had been at its beginning. Until about 1720, European craftsmen sought to reproduce exactly the figures, the architectural settings, and the stylized vegetable forms of the imported lacquers. Then, in keeping with the playful spirit of the rococo, they modified these designs by introducing European figures, exotic animals such as monkeys, draperies and arabesques, and cartouches and ribbon compositions. Along with this transformation, in place of the conventional black and gold of the Oriental products, came a wide choice of background colours, including scarlet, yellow, white, blue, and green, sometimes flecked with gold. Particularly in Venice, where craftsmen followed the rules of a treatise by Filippo Buonanni (1722), a great originality was achieved by the informal spacing of bouquets of flowers around gracefully posed figures set against delicate hues of yellow and bluish green.

The uses of lacquer in Europe reflect the changing tastes of the 17th and 18th centuries. In the late 17th century it was principally employed for decorating the cases of cabinets set upon carved Baroque bases or for ornamenting leather wall coverings. In the 18th century, bookcase desks, tall clocks, and tea tables became the most fashionable articles of lacquered furniture in England and Germany, while in France and Italy the chest of drawers and corner cabinet were preferred. Whole sets of furniture with Orientalizing lacquer decoration were made in England by Giles Grendey (Figure 192) and other cabinetmakers of the Chippendale era (1754–68), many of whose masterpieces are housed in the Victoria and Albert Museum, London. Rooms with lacquered walls have survived at the palaces of Nymphenburg outside Munich, Bamberg, and other German cities. Small lacquered boxes called *Boîte*

By courtesy of the Victoria and Albert Museum, London

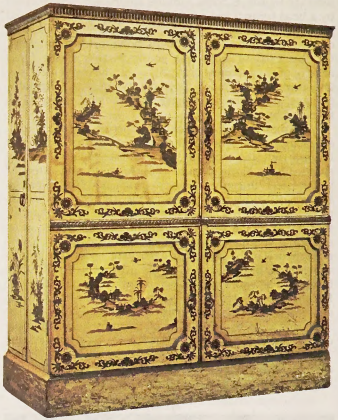


Figure 192: Lacquered armoire by Giles Grendey (1693–1780). In the Victoria and Albert Museum, London. 1.4 m × 1.68 m × 62 cm.

Reproduced by permission of the Trustees of the Wallace Collection, London



Figure 191: Lacquered commode attributed to René Dubois, 18th century. In the Wallace Collection, London. 1.57 m × 91 cm.

de Spa became a specialty of that Belgian town and the nearby centres of Liège and Aachen, where a member of the Dagly family was active.

In the early 19th century the solemn grandeur of the international Classical style with its insistence on plain wood or gilt surfaces almost eliminated the taste for lacquered furnishings. It survived in Victorian times in the vogue for painted tin or toleware (as in Pontypool, England) and the decoration of small tables and chairs of papier-maché inlaid with mother-of-pearl. The making of this furniture, dominated by the London firm of Jennens and Bettridge, rapidly declined after reaching the height of its popularity about the time of the Great Exhibition held in London in 1851. In the period 1925–30, two European sculptors, the Belgian Marcel Wolfers and the Franco-Swiss Jean Dunand, successfully approximated the true Oriental lacquer techniques in a few art objects of clay, wood, and bronze. By this time, however, the making of lacquer surfaces had become a part of the chemical industry, which today produces effects of depth and hard translucence surpassing the finest products of the European lacquer masters of the past. (R.C.Sm.)

MOSAIC

Mosaic is the art of decorating a surface with designs made up of closely set, usually variously coloured, small pieces of material such as stone, mineral, glass, tile, or shell. Unlike inlay, in which the pieces to be applied are set into a surface that has been hollowed out to receive the design, mosaic pieces are applied onto a surface, which has been prepared with an adhesive. Mosaic also differs from inlay in the size of its components. Mosaic pieces are anonymous fractions of the design and rarely have the dimensions of pieces for intarsia work (fitted inlay usually of wood), whose function is often the rendering of a whole portion of a figure or pattern. Once disassembled, a mosaic cannot be reassembled on the basis of the form of its individual pieces.

Technical insight is the key to both the creation and the appreciation of mosaic, and the technical aspects of the art therefore require special emphasis in the present section, both in those portions devoted to materials and methods and in the historical survey of the major periods and centres of mosaic activity. Stylistic, religious, and cultural aspects of mosaic, its role in Western art, and its appearance in other cultures are also dealt with in order to complete the picture of an art form that appears in widely separated places and at different times in history, but only in one place—Byzantium—and at one time—4th to 14th centuries—rose to become the leading pictorial art.

Principles of design

Between mosaic and painting, the art with which it has most in common, there has been a reciprocal influence of varying intensity. In colour and style the earliest known Greek figurative mosaics with representational motifs, which date from the end of the 5th century BC, resemble contemporary vase painting, especially in their outline drawing and use of very dark backgrounds. The mosaics of the 4th century tended to copy the style of wall paintings, as is seen in the introduction of a strip of ground below the figures, of shading, and of other manifestations of a preoccupation with pictorial space. In late Hellenistic times there evolved a type of mosaic whose colour gradations and delicate shading techniques suggest an attempt at exact reproduction of qualities typical of the art of painting.

Interaction of painting and mosaic

In Roman imperial times, however, an important change occurred when mosaic gradually developed its own aesthetic laws. Still basically a medium used for floors, its new rules of composition were governed by a conception of perspective and choice of viewpoint different from those of wall decoration. Equally important was a simplification of form brought about by the demand for more expeditious production methods. In the same period, the increasing use of more strongly coloured materials also stimulated the growing autonomy of mosaic from painting. As a means of covering walls and vaults, mosaic finally realized its full potentialities for striking and suggestive distance effects, which surpass those of painting.

The general trend towards stylization—that is, reduction to two-dimensionality—in late antique Roman painting (3rd and 4th centuries AD) may have been stimulated by experimentation with colour in mosaic and particularly by the elimination of many middle tones for the sake of greater brilliance. The central role played at that time by mosaic in church decoration, for which it is particularly well suited, encourages the assumption that the roles had shifted and painting had come under its influence. The strong, sinuous outlines and the absence of shading that came to characterize painting during certain periods of Byzantine and western European art of the Middle Ages may have originated in mosaic technique and use of materials. It is notable, however, that from the Renaissance to the 20th century, mosaic was again wholly dependent on painting and its particular forms of illusionism.

In modern mosaic practice, the main tendency is to build on the unique and inimitable qualities of the medium. Although not a few of the works created in the 20th century reveal the influence of painting, figurative or abstract, the art has gone a long way towards self-realization. By and large the modern mosaic maker shares with his medieval predecessor the conviction that there are functions to which the materials of mosaic lend themselves with particular appropriateness.

Materials

In antiquity, mosaics first were made of uncut pebbles of uniform size. The Greeks, who elevated the pebble mosaic to an art of great refinement, also invented the so-called tessera technique. Tesserae (Latin for "cubes" or "dice") are pieces that have been cut to a triangular, square, or other regular shape so that they will fit closely into the grid of cubes that make up the mosaic surface. The invention of tesserae must have been motivated by a desire to obtain densely set mosaic pictures which could match, in pavements, the splendour of contemporary achievements in painting.

Invention of the tessera technique

Tesserae vary considerably in size. The finest mosaics of antiquity were made of tesserae cut from glass threads or splinters of stone; ordinary floor decorations consisted of cubes about one centimetre square. Medieval works often display a differentiation in tessera size based on function: areas requiring a wealth of details, faces and hands, for instance, are sometimes set with tesserae smaller than average, while dress and jewelery are occasionally set with very large, single pieces.

As long as mosaic was a technique for the making of floors, the main requisite of its materials, besides their colour, was their resistance to wear.

STONE

Stone, therefore, was long dominant, and throughout antiquity the natural colours of stone provided the basic range of tints at the artist's disposal; they put their mark not only on the earliest Greek works but continued to determine colour schemes far into Roman times. Stone continued to be used in Christian monumental decorations but on a more limited scale and for special effects. In Byzantine mosaics, faces, hands and feet, for example, were set with stone, while cubes of marble, often of coarse crystals, were used to depict woollen garments. Stone was also used for background details (rocks, buildings), probably to bring about particular illusions. Though marble and limestone were ordinarily preferred, in a period when Roman mosaic cultivated a black and white technique, black basalt was widely employed. Marble cubes painted red, probably to substitute for red glass, have been found in many Byzantine mosaics, in 9th-century works at Istanbul, for example.

Because its granular, nonpolished surface is often preferred to the hard brilliance of other materials, stone is also widely used in modern mosaics. At the University of Mexico in Mexico City, for example, the mosaics covering the exterior of the library by Juan O'Gorman (1951–53) and the exterior of the stadium by Diego Rivera (1957) are made with natural stone (Figure 193).

GLASS

Glass, which first appeared among the materials of mosaic in the Hellenistic period (3rd–1st century BC), brought unlimited colour possibilities to the art. In floors, however, it had to be used sparingly because of its brittleness. In floors, glass tesserae were used for the strongest hues of red, green, and blue, while softer tints were rendered with coloured stone. With the development of wall mosaic, glass largely took over the functions of stone, producing tints of unsurpassed intensity and leading to a continuing search for new coloristic effects.

With little knowledge of the laws of optics but with immense practical experience, mosaic makers of the Early Christian period gave the art a completely new direction with the exploitation of gold and silver glass tesserae. Like

Gold and silver tesserae



Figure 193: Stone mosaic, detail from the stadium at the University of Mexico, Mexico City; by Diego Rivera, 1957.
Harrison Forman

a mirror, the glass from which this kind of tesserae was made had a metal foil applied or, better, encased in it. The metal was gold leaf or, for the "silver," probably tin. These pieces of mirror glass gave golden or white reflections of high intensity and could be used to depict objects of precious metal or to heighten the effect of other colours; but, above all, it was used as a means of rendering the light emanating from God.

Gold tesserae were first used by the Romans, in both floor and vault decoration of late antiquity. Initially, their role was simply to give a golden effect. Gold tesserae, for example, were employed to depict a golden wreath in a floor mosaic at Antioch (*c.* AD 300) and gold vessels in some of the vault mosaics in Sta. Costanza in Rome (Figure 194). Later, when this use of gold for imitation purposes had become more refined, some spectacular effects were produced in the depiction of garments. The Good Shepherd in the Mausoleum of Galla Placidia at Ravenna (*c.* AD 450) is dressed in golden robes of densely set gold cubes shaded with stripes of light-yellow tesserae. The female saints in S. Apollinare Nuovo (*c.* AD 550–570) in the same town wear costumes set with green glass cubes among which appear both patterns and large fields of gold tesserae, producing a striking similarity to rich silk brocade. Silver was used in a similar way. Christ scenes in S. Apollinare Nuovo (AD 500–526) employ silver tesserae in the drawn sword of Peter in the betrayal, no doubt an imitation of steel. Silver tesserae are also found in the silver jug and basin in the scene of Pilate washing his hands.

Gold cubes were distributed among the ordinary tesserae to add to the shimmer of light in ornaments and background details. To avoid an uneven gleam in the surface, the mirror effect was often moderated by setting the gold tesserae in reverse, so that the visible part of the cube is the side with the thickest sheet of glass covering the gold leaf. In the now-lost mosaics of the Church of the Dormition in Nicaea, a scholar observed another exquisite effect, which he called dark gold, created by cubes from which some of the gold leaf had been chipped off, for example, in the frontal part of Mary's golden footstool (7th or 8th century AD).

An early instance of the use of gold for depicting light emanating from God is in a representation of Christ-Helios (Christ as the Sun God) in a 3rd-century mausoleum under St. Peter's at Rome. Here, a few gold tesserae are seen in the rays coming from Christ's head. The halo of gold, a feature so common in Christian art that religious pictures without it can hardly be imagined, developed in mosaic art in the 4th century AD. The gold background, signifying divine light, probably originated in Roman mosaic art, but the first preserved instances date from the advanced 4th century. The cupola mosaic of Áyios Geórgios at Thessaloniki (*c.* 400), for example, has a background of gold. In Italian mosaics of the 5th century, other types of background, such as a dark-blue ground or a more naturalistic landscape setting, were dominant. Only at the beginning of the 6th century did the gold background become the rule.

In addition to this massive predilection for gold, the Christian East began to use silver to depict the symbolic light emanating from Christ. First, it was used for the entire disc of his halo, later only for the cross arms. The archangels were the only figures besides Christ for whom the silver halo was used. The light of God, appearing as rays from above in scenes of the Annunciation, Nativity,

SCALA, New York

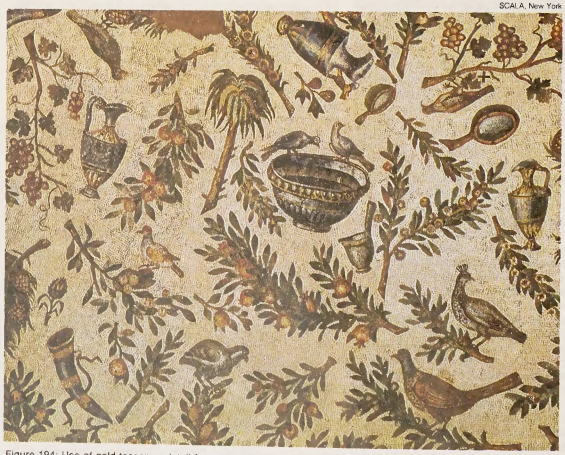


Figure 194: Use of gold tesserae, detail from the Early Christian vault mosaics of Sta. Costanza, Rome, *c.* 337–354 AD.

Baptism, and Transfiguration, was also depicted with silver tesserae. Finally, silver and gold were used together in Byzantine representations of the infant Jesus whose golden robes are highlighted with silver cubes (the apse and south vestibule of Hagia Sophia, Istanbul; both 9th century).

OTHER MATERIALS

In Christian mosaics, tesserae of mother-of-pearl or coarse-grained marble cut to round or oblong shapes were used to depict pearl. Though pieces of semiprecious stones were among the mosaic materials of antiquity, their use was rarely dictated by the wish for particular sumptuous effects. Reduced to common tessera size, bits of this strongly coloured material served as part of the general colour scheme of the mosaic pictures. Objects like those of the pre-Columbian American Indian cultures, in which, because of its exquisite materials, such as turquoise and garnet, mosaic attained the status of jewelry have not been found in Western art.

Ceramic tesserae

Among the materials that have played and continued to play a role in the production of mosaic, ceramic is the most versatile. Terra-cotta "threads" were used in Greek mosaics as contours, and tesserae of the same material were frequently used by the Byzantines for the depiction of red objects and garments. Today, glazed or unglazed ceramic is used and is one of the strongest competitors with glass and stone. Ceramic tesserae are cut from tiles or, like much modern glass mosaic material such as pressed glass, come prefabricated. Prefabricated tesserae have the advantage of a very uniform and smooth surface which harmonizes with glass, steel, and other new building materials.

Techniques

The most commonly used adhesive for mosaics was mortar, the function of which in the 20th century has been largely taken over by modern, tougher kinds of cements or glue. In Roman floors, two to three layers of mortar preceded the setting bed that was to carry a tesserae facing. The first layer rested on a thick foundation of stone that prevented settling of the mortar bed and the formation of cracks. For wall mosaics the preparation was equally painstaking, and in many cases an application of a waterproofing of resin or tar preceded the laying of the mortar. There then followed two layers of coarse, roughened mortar, the stability of which was often improved by large nails that had been driven into the joints of the wall before the work of laying started. A third and final layer was of fine consistency and frequently, like the mortar for floor mosaics, contained powdered marble and binding elements such as pounded brick.

As in fresco painting (technique of using water-suspended pigments in a moist plaster surface), the setting bed was applied in patches never larger than were needed for one day's work. In a frescoed surface, the breaks between the different stages of the work can easily be detected; they are harder to discover in mosaic.

Sketches and underpainting

Numerous underpaintings discovered in wall mosaics indicate that sketches, often detailed and with the main colours suggested, were executed on the setting bed to serve as guides for the disposition of the tesserae. Similar procedures are thought to have been part of the technique of floor mosaic. In church mosaics, rough preliminary sketches have been found on layers underneath the setting bed and, in a few instances, even on the brick wall itself. This kind of preparatory sketch, for which there are parallels in wall painting, suggests that the artist was trying out the overall scheme of the decoration before making a more detailed sketch on the setting bed.

Instead of laying the tesserae one by one directly onto the mortar, another method was sometimes used. In Pompeii many of the so-called *emblêmata* (central panels of floors), which were made up of smaller than average tesserae and were often of very high artistic quality, appear to have been preset on trays of stone or terra-cotta which were then embedded in the mortar of the floor. The surrounding mosaic area was then set according to the ordinary, direct method. Although the direct method was used for wall mosaics during the Middle Ages, there are signs in at least one medieval monument of a partial use of the prefabrication—or "indirect"—method: in the cupola mosaic of the church of Áyios Geórgios, Thessaloníki (*c.* 400), the heads of the saints seem to have been inserted in the mortar in one piece. The indirect method is the one most used in the 20th century. In the workshop, the mosaic is first set in reverse with glue on paper or cloth and then applied to the floor or wall. The technique permits preassembling of mosaics intended even for curved surfaces, cupolas, or apses. It has been hypothesized that behind the enormous output of floor mosaics in the Roman era lay similar production methods which had developed out of the tray procedure described above. The introduction of wall mosaics led to experimentation with the spacing and angling of tesserae. The solidity of floor mosaics depended on a close-set texture, but in wall mosaics, in which the element of wear was no longer relevant, the organization of the surface could become looser. For several centuries, a very wide spacing of the tesserae was cultivated, and the placing of cubes at irregular angles was regarded as important to the over-all effect of wall mosaics. These tendencies reached the extreme in the 7th and 8th centuries, in mosaics of the chapel of S. Venanzio in the Lateran Baptistery, Rome (Figure 195), and in the fragments of the

SCALA—Art Resource

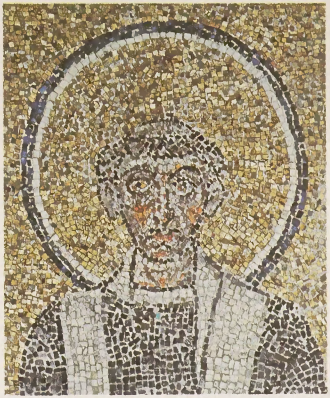


Figure 195: Wide-spaced tesserae set at irregular angles in the head of a saint, detail from mosaics in the Chapel of S. Venanzio, Lateran Baptistery, Rome, *c.* 640 AD.

decoration of Pope John VII (AD 705–707) in the old St. Peter's in the Vatican. Later periods preferred a somewhat closer setting, but the irregular surface continued to be in fashion for most of the Middle Ages.

Tilting of tesserae

The tilting of tesserae became an art in itself. In 6th-century Byzantine mosaics, there evolved a new technique whereby gold and silver tesserae were set at extremely sharp angles to enhance reflection. By pointing their mirror ends downward in the direction of the onlooker it was possible to secure maximum light effect. In Hagia Sophia at Istanbul, the enormous gold areas in the wall mosaics of the emperor Justinian are set with cubes tilted this way. In one particularly dark corner, the tesserae are not only tilted downward but are also turned slightly sideways to catch the light from a nearby window. A similar technique, based on a high degree of tilting of the gold tesserae in unlit areas, can be observed in the mosaics of the Dome of the Rock in Jerusalem (*c.* AD 690).

Haloes set with tilted cubes that bring out the circle of light surrounding the heads of holy figures became common in Byzantine mosaics of the 6th to 7th centuries,

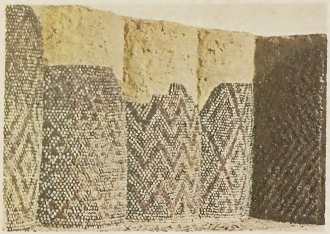


Figure 196: Columns decorated by the Sumerians in a mosaic-like technique with polychrome terra-cotta cones, from Uruk, Mesopotamia, early 3rd millennium BC. In the Staatliche Museen zu Berlin.
By courtesy of the Staatliche Museen zu Berlin

as is seen in the mosaic panels dating from this period in the church of Ayios Dhimítrios, Thessaloníki. Striking examples of such haloes are also found among mosaics that were put up in Hagia Sophia in Istanbul in the 9th century, above all in a panel with the kneeling emperor (Leo VI?).

Effects such as those described above are unthinkable without the accumulated experience of the craftsman-artist. In the 20th century, mosaic increasingly has become an art divided between the inventor who furnishes the design and the worker who executes it. It may be that the dry character of many modern mosaics can be ascribed to the fact that the artist no longer puts his thumb on every tessera.

Periods and centres of activity

Among the cultures of the ancient Near East there is one remarkable occurrence of a mosaic-like technique: the exteriors of some large architectural structures, dating from the 3rd millennium BC, at Uruk (Erech) in Mesopotamia, are decorated with long terra-cotta cones imbedded in the wall surface (Figure 196). The blunt, outer ends of the cones, coloured in red, black, and white, form patterns consisting of zigzag lines, lozenges, and other geometrical motifs. This revetment was decorative as well as functional, for the cones shielded the core of sun-dried bricks from rain and wind. The technique, however, died out and seems to have had no influence on the later development of mosaic.

Terra-cotta cone mosaics

In western Asia Minor are preserved the earliest examples of the surface-covering technique that lead to mosaic in the present sense of the word. In the town of Gordium near modern Ankara in Turkey, houses have been uncovered with floors made of pebbles set in a primitive mortar. In some of these floors (dated to the 8th century BC), rows of light pebbles form awkward geometrical figures against a background of darker stones. These rudimentary elements of decoration introduced into a crude form of pavement laying spurred artistic imagination and set in motion a process that was to bring spectacular results.

ANCIENT GREEK AND HELLENISTIC MOSAICS

Three main phases can be determined in the development of mosaic art in antiquity. The first, chiefly a Greek matter, involved the gradual perfecting of the pebble medium. The second, which saw the invention and spreading of the tessera technique, took place partly in the Hellenistic Greek world and partly on Roman soil. The third, largely a Roman phenomenon, was characterized by the popularization of mosaic and the application of the medium to new functions. By a process of diffusion, the taste in floor decoration documented at Gordion spread through the Greek-speaking world of the Mediterranean. Its first full flourishing seems to have occurred in late Classical times. Pebble mosaics are found as far west as Sicily (Motya, Morgantina) and, in the east, in the Greek colonies on the Crimea (Cherson). They are preserved in large number at only two sites, Olinthos and Pella, in Macedonian northern Greece.

In the town of Olinthos there are floor mosaics, with elaborate figures and complicated patterns, which were part of the new city culture that developed in the 5th century BC. The Olinthos mosaics also reveal that picture making with light and dark pebbles had by then evolved into an intricate art. Against a ground set with black or blue-black stones stand figures or patterns set with white or slightly tinted ones. Pebble size had become fairly uniform, with diameters from one to two centimetres, but particularly intricate areas, like faces, are set with smaller stones. Very small black pebbles serve as outlines. Although the mortar between them is visible, the pebbles are set close enough so that the pictures appear with regular, not too broken outlines and with some emphasis on detail.

Floors in houses at Pella, dating from the 4th century BC, demonstrate a significant later development of the pebble technique (Figure 197). Floor mosaics then openly vied with wall painting in the rendering of space and realistic detail. This was made possible by the introduction of new materials that eliminated the shortcomings of the ordinary pebble medium. Among the basic changes was an increase in the range of colours. When the demand for particular tints could not be met with pebbles of natural colours, "artificial pebbles" were made and are found in several of

Dimitri



Figure 197: Greek pebble mosaic, detail from The Lion Hunt, from Pella, Macedonia, c. 300 BC.

the floors at Pella. These "pebbles" have been painted in the required tones—mostly strong green and red—and, to protect the film of paint, have a depression sunk in the middle. The new trend also called for smaller pebbles to permit pictures to be set more closely. To obtain precise delineation of limbs and features, outlines made not with pebbles but with long strips of terra-cotta or lead wire were employed. Pictures made in this technique reflect a taste for heroic hunt scenes and fights with wild animals, themes inspired from court art glorifying the ruler.

The next innovation came at the end of the 4th or the beginning of the 3rd century BC, when the introduction of new principles led to the abandonment of the pebble technique. Cut mosaic pieces permitted the nearly complete elimination of the disturbing effects of visible mortar patches, and new materials, above all glass, offered a vast new range of colours. The acceptance of the new methods

and materials seems, however, to have come about slowly. In Alexandria (Greco-Roman Museum) there is a mosaic (depicting Erotes fighting a stag and, in the outer border, a frieze of animals) in which cut, triangular tesserae are used together with both pebbles and lead wire. In somewhat later Alexandrian mosaics made with tesserae (late 3rd/early 2nd century BC) lead threads are still in use; for example, in a panel signed by the artist Sophilos and depicting a personification of Alexandria, that is the earliest known example of miniature mosaic work (called *opus vermiculatum*, meaning "wormlike work" because of the close-set, undulating rows of small tesserae).

The Pergamum School

Pergamum, another centre of the Hellenistic world, was particularly famous for its school of mosaics. According to the ancient Roman historian Pliny the Younger, Sosos, one of the most renowned mosaic artists of antiquity, worked in this city. None of his works survives but, thanks to Roman copies, the intentions that underlay his art can be judged. Pliny listed as his most celebrated works a representation of drinking doves and a clever imitation of the "unswept floor" of a banquet room (*asarōtos oikos*). The copies tell of Sosos' phenomenal ability to create trompe l'oeil ("fool-the-eye") effects through a shading and colouring that seems to bring the objects out in full plasticity on the ground on which they are depicted. To call the work merely an imitation of painting may be incorrect. The intense colours and the smooth texture permitted by the new setting technique paved the way for illusionistic effects that went beyond those achieved by painting.

ROMAN MOSAICS

Eager to adopt the artistic culture of the Hellenized eastern Mediterranean, the Romans introduced mosaic in this exquisite form in both their domestic architecture and their places of worship. Pompeii has yielded a host of *opus vermiculatum* works datable to the 2nd/1st century BC. Among these the most famous is the Battle of Issus, found in the Casa del Fauno in 1831 (Figure 198). This is the largest of all known works, measuring about 11.22 by 19.42 feet (3.42 by 5.92 metres), in the miniature mosaic technique. This mosaic (which probably copies a work of painting, perhaps a famous picture by Philoxenus of Eretria) and other Pompeiian panels of similar quality are supposed to have been executed by Greek artists, who carried on in the tradition established at Alexandria and Pergamum.

SCALA—Art Resource



Figure 198: "Battle of Alexander and Darius at Issus," detail of the Roman mosaic done in the *opus vermiculatum* technique, from the Casa del Fauno, Pompeii, late 2nd century BC. In the Museo Archeologico Nazionale, Naples.

The Romans transformed mosaic from an exclusive art to a common decorative medium. Some of the earliest examples of this new type of floor are in the late republican (2nd century BC) houses at Delos. For rooms of secondary importance and often for floors surrounding the finely designed and executed central *emblēmata* (a featured picture or ornamental motif) in the most important rooms, the Romans developed a simpler, less artistic kind of mosaic. The floors are set with fairly large tesserae with a limited range of colours, some tending toward monochrome (black-and-white). The decorative designs and motifs are also simple and uncomplicated.

SCALA—Art Resource



Figure 199: Roman monochrome floor mosaic in the Portico delle Corporazioni, Ostia, Italy, 3rd century AD.

Roman floor mosaics

This new trend in mosaic floors was probably stimulated by new and functional ways of thinking about the role of floors in architecture. To the practical Romans it may have seemed illogical that floors destined for rough wear should bear delicate pictures. Moreover, the demand for large-scale mosaic making brought about by the colossal urban expansion in the 1st century AD made the development of quicker and simpler techniques imperative. The aim of the Romans seems to have been to create a style, technique, and form of composition that would be simple and functional. Competition with painting in illusionistic and coloristic refinement was therefore abandoned; *emblēmata* gave way to decorative elements distributed over the floor in one large overall pattern or to figure compositions taking the full floor plane; and polychrome gave way to monochrome mosaics (which may have been easier to produce). Enormous floors in the baths and in the courtyards of warehouses (1st to 3rd century AD) at Ostia, Rome's port at the mouth of the Tiber, are the best preserved examples of the monochrome style (Figure 199).

The expressionist Roman style, which flourished in Italy, penetrated into the former Greek cities in the eastern part of the empire, but polychromy and types of composition based on the framed picture persisted with especial tenacity due to strong local Hellenistic traditions. A splendid series of *emblēmata* (2nd century) with mythological representations, allegories, and scenes from the theatre have been uncovered at Antioch in southern Turkey. They prove the existence of a school there of mosaicists of particular brilliance. Recent research has pointed to the African provinces as the site of another, highly active school with a taste for larger, dramatic compositions. Influence from these areas may have been responsible for the renewed opulence, represented by a vivid polychrome pictorial mosaic, which reappeared in Roman art in late antiquity. Outstanding examples of this renewal are the mosaics in the Roman villa of Casale (*c.* AD 300) near Piazza Armerina, Sicily. The mosaic decoration of this vast palace complex culminates in the gallery of the Large Hunt, which contains a scene of animal hunting

and fighting covering an area of 3,200 square feet (300 square metres).

It is generally agreed that in the course of the 3rd century the status of mosaic was radically altered. Already in Hellenistic times the medium had been employed for other ends than floor covering and had become part of the embellishment of the fantastic garden architecture of which the rulers of the period seem to have been particularly fond. Reflections of this tradition in the 1st century AD are the mosaic-covered fountains in the mansions at Pompeii and Herculaneum and mosaic panels and niches in rustic banquet halls and artificial grottoes at The Golden House of Nero in Rome and his villa at Anzio. Mosaic fragments and imprints of tesserae in the vaults of baths and buildings of similar size demonstrate that mosaic gradually was introduced into new fields. Equally important is the evidence that mosaic was used to depict sacred images. On some of the monochrome floors at Ostia are scenes pertaining to animal sacrifice and to the cult of the dead. Three monuments of the 3rd century inform of another new practice introduced at this time, that of putting mosaic pictures of religious importance on walls: a niche mosaic with the god Silvanus from a temple of Mithra at Ostia; a Christian wall and vault mosaic depicting Christ as Helios, the Sun God, in a mausoleum under St. Peter's, Rome; and a decoration, now lost but recorded in a 17th-century drawing, of a chapel for the Lupercalian worship at Rome. It has been pointed out by modern scholarship that the new role gradually assumed by mosaic must be related to the corresponding decline in interest in three-dimensional representation. The cultic mosaic took over the function of the cult statue, mosaic being that two-dimensional medium which was considered most capable of convincingly expressing religious ideas in visual form.

Cultic mosaics

EARLY CHRISTIAN MOSAICS

Present-day insight into the crucial, early phase of this part of the history of mosaic is limited because of the loss of nearly everything that was made in the field during the first half of the 4th century. Nevertheless, as indicated above, it seems certain that wall mosaics had come into use in Roman art well before Emperor Constantine's edict of toleration of the Christian faith in AD 313. Considered to be among the earliest Christian wall mosaics in Rome are those in the church of Sta. Costanza built about AD 320–330 as a mausoleum for Constantine's daughter. The content of the pictures is almost completely Dionysiac and pagan, but a series of small format scenes from the Old and New Testaments were included among the non-Christian pictorial elements of the decoration. Obviously an independent Christian pictorial program for buildings of Sta. Costanza's size and complexity had not yet been developed; and, probably in lieu of that, a Dionysiac program had been chosen because its many allusions to the symbolism of wine lent themselves to a Christian interpretation.

Other monuments of the 4th century bear similar marks of transition. Floor mosaics in the cathedral complex at Aquileia demonstrate that the church before and immediately after Constantine's edict of tolerance of the Christian faith in AD 313 adhered to the late antique tradition of placing religious pictures in pavements. In the earliest group of Aquileia mosaics (*c.* AD 300) objects and animals symbolize the Good Shepherd, while the later group (second decade of 4th century) contains scenes from the story of Jonah, symbolic animals, such as the deer and the lamb, and a representation of the bread and the wine. Before long, pictures of this character were banished from floors, and simpler and more general symbols took their place.

The latest of the preserved transitional works, a decorated cupola in a mausoleum, possibly imperial, at Centcelles (now Constantí, Tarragona), Spain, seems to have been made not long after AD 350. This very fragmentary decoration has yielded important information about a stage of increasing mastery in the handling of the medium. The scenes from the Old and the New Testament are presented with greater self-confidence and occupy a full, broad zone in the lower part of the cupola. Yet, below it is a stag hunt, rich in symbolic content but adhering closely to the patterns of profane floor mosaics. Stone tesserae dominate in the lower zones, but glass cubes are found in large quantities in the upper. Glass, with its stronger colours, was doubtlessly concentrated in this area intentionally. The zenith of the cupola, weak in lighting and distant from the spectator, needed tesserae of strong reflecting power to make it possible to read its decoration.

A series of large, in part well-preserved mosaics make it possible to follow the progress of the art in 5th-century Italy. Ravenna and Rome have several important works, while Naples and Milan have preserved enough to suggest that workshops of high artistic standard must have existed in many of the large cities of the peninsula. In these works, the tendency to clarify and even underline the content of religious pictures with the help of colour is brought to its full peak. The swing towards a greater employment of glass reached a point at which the mosaics are almost entirely made of this material. In what must be regarded as a late but vigorous revival of the painterly illusionism of antiquity, there is an audacious blending of colours. Among the high points of this trend are the flaming visages of angels in Sta. Maria Maggiore, Rome (*c.* AD 432–440) and the spiritualized physiognomies of St. Bartholomew and his fellow apostles in the Baptistery of the Orthodox, Ravenna (*c.* 450). But the designer's mastery and sophistication are nowhere more overwhelmingly illustrated than in the glowing interior of the so-called Mausoleum of Galla Placidia (*c.* 450) at Ravenna, with its blue star-filled mosaic dome, and in the decoration of the Naples Baptistery of S. Giovanni in Fonte (5th century), with its hypnotizing glimmer.

Development of an iconographic tradition

A Christian language of pictures (iconography) was now developed, and its grammar worked out. In cupolas, the centre tended to be reserved for depictions of Christ or the cross. In apses there was a trend toward static and symbolical representation of holy figures and a reduction of detail. On the walls of the nave of basilicas were scenes from the Old or the New Testament or both. The largely intact decoration of the church of Sta. Maria Maggiore, Rome, throws some light on the principles involved.

Old Testament scenes are distributed on the side walls of the nave in panels measuring about six by six feet (1.9 by 1.9 metres). There is one panel below each of the basilica's large windows. The pilasters (columns projecting shallowly from the surface of the wall) between the windows (restored but repeating the original disposition) serve as outer frames for these panels. Before the restoration of the church in the 16th century there were also inner frames, made of stucco; in addition, each panel was adorned with a small pediment (triangular gable) of the same material and thus appeared as if enshrined by a small aedicula (a pedimented niche). The classical rules governing the relation between the architecture of a building and its decoration may be expected to leave their mark on Christian mosaic art for a long time.

BYZANTINE MOSAICS

Early Byzantine mosaics. Mosaics made in Ravenna for the Ostrogoth king Theodoric (AD 493–526) are the first full manifestations of Byzantine art in the West. As seen in two of the foremost works from his time, the Baptistery of the Arians and the church of S. Apollinare Nuovo, the gold background now dominates. Accompanying it was silver, a novelty among the mosaics of Italy. In S. Apollinare Nuovo, the faces and hands in several of the Christ scenes are set not with glass tesserae but with cubes of stone. Stylistically, these mosaics are characterized by more static figures and less depth and plasticity than in those of the 5th century.

Developed integration of architecture and mosaic decoration

Another remarkable element is the movement, now fully developed, toward an integration of architecture and mosaic decoration. This is most clearly seen in the basilica. In the church of S. Apollinare Nuovo, the mosaics are no longer inserted panels but form a continuous "skin" that covers every inch of the wall. The size of the windows and even their number have been reduced, apparently to provide more wall space for pictures. Figures have grown in size, to match the dimensions of architectural components, and they seem to have taken the place of pilasters

in the articulation of the room. This trend is given truly monumental expression in the choir of the church of S. Vitale at Ravenna, dating from *c.* AD 526–548. A profusion of decorative elements is spread like tapestry over the walls and vault, the panels of the emperor Justinian and his consort Theodora near the apse embodying the new spirit in their colour-laden pageantry.

In the East, the circular church of Áyios Geórgios at Thessaloníki, Greece, shows Byzantine mosaic at its earliest flourishing (*c.* AD 400). Its partly preserved mosaics display a disposition related to that of the Baptistery of the Orthodox at Ravenna, with a lower zone containing Paradisiac architecture; above this a zone with standing and walking figures and in the centre of the cupola a medallion with a figure of Christ. A uniform gold background dominates the two lower zones, at a time well before it had come into general use in the West. Silver is found in profusion, used for the background in the central medallion as well as a means to enhance the radiation of light from all parts of the mosaic. In the figures of saints the material for faces and hands is chiefly natural stone, its gentle gradations contrasting spectacularly with the violent juxtapositions of coloured glass tesserae of the hair and the garments.

Effects of this kind, which are the hallmark of Byzantine mosaic technique, seem to derive from intentions wholly dissimilar to those which had determined the development of early Christian mosaic art in the West, where the illusionism of Greco-Roman art persisted for a long time. In the great Ravenna mosaics of the 5th century, pictures illustrating the narrative of the Bible or expounding the dogmas of religion were still done in the painterly style of Roman mosaics and wall painting. During the same period, mosaic art of the Eastern Empire, having abandoned conventional illustration, was boldly exploring the way that lay open, in mosaic art, toward a new kind of imagery.

In the mosaics of the 6th century are found the earliest refinement introduced by the Byzantines to enhance the brilliance of gold tesserae. This refinement, already described, involved setting gold cubes at oblique angles to direct their reflections toward the viewer. Used in haloes, the tesserae, obliquely set, convey to the holy figures a miraculous aura of light. The visages of the saints, with their dull stone surfaces and hues reminiscent of actual human skin, add a touch of mysterious reality to this theatre of effects.

Rene Percheron—J.P. Ziolo



Figure 200: Gold tesserae reflecting light to the viewer; from the Byzantine votive mosaic showing St. Demetrios between two donors, from Áyios Dhimítrios, Thessaloníki, Greece, 7th century.

Splendid mosaics from many parts of the eastern Mediterranean testify to the continuous cultivation and improvement of these effects. In the city of Thessaloníki the mosaics in the churches of Hosios David (5th century AD) and Ayios Dhimítrios (6th and 7th centuries) exemplify the trend (Figure 200), which is also expressed in apse decorations preserved at Cyprus (church of the Panagia Angeloktistós, at Kiti, and of the Panayía Kanakaria near Lythrangome; both 6th century) and in the Monastery of St. Catherine, Sinai Desert, founded by Justinian.

Apart from the gold ground, which had considerable impact, the technical subtleties essential to these mosaics met very little response outside Byzantium. When Byzantine artisans operated in foreign territory, they brought their particular techniques with them. Again and again the impact of this tradition was felt in the West, though, at its purest, mostly as short-lived episodes. To judge from a few surviving fragments, mosaics executed under Pope John VII (AD 705–707) in a chapel in St. Peter's, Rome, might have been the work of artisans summoned from Byzantium. Technical and stylistical features demonstrate that the mosaics executed under the earliest Muslim rulers, in the Dome of the Rock at Jerusalem (*c.* AD 690) and in the Great Mosque at Damascus (*c.* AD 715), are certainly the work of specialists called from Byzantium. Sources testify that even the mosaics in the mosque at Córdoba, Spain (AD 965), were made by Greek craftsmen.

The floor mosaics in the great palace of the Byzantine emperors at Istanbul—with their pastoral scenes, fights with wild animals, and figure groups taken from pagan mythology—testify to an undercurrent of classical taste in Constantinople. The date, which according to archaeological evidence must be placed as late as around the year AD 600, demonstrates the tenacity of that taste in the midst of the Christian milieu of the Byzantine metropolis.

Middle Byzantine mosaics. Scholars have been concerned to discover how Iconoclasm, the dispute concerning images during the 8th and 9th centuries, may have influenced the course of Byzantine art. In some respects, at least, mosaic reflects very little change. The main source of knowledge about the state of mosaic in the time shortly after the end of Iconoclasm is Hagia Sophia at Istanbul. Parts of the redecoration that the church underwent in the last half of the 9th century have been uncovered in recent times. In their colour and technique these show a continuation of the early Byzantine tradition: the preference for rather strong, clear tints, and the effects created by such techniques as the tilting of tesserae and the turning of gold cubes. The preoccupation with light seems stronger than ever: in badly lit places in the vestibule and gallery, the gold ground displays a high percentage of silver cubes among the gold ones to add to the sparkle. Stylistically, new ground had been broken. Particularly in faces, the tesserae are set in wavy lines which break up the modelling in bandlike configurations. Linearism (the expression of form in terms of line rather than colour and tone) had taken a great step forward.

Mosaics of Hagia Sophia

In the arrangement and distribution of pictures new features are visible. In the apse of Hagia Sophia, the Virgin with the Child sits surrounded by a vast expanse of gold. She is one of the first of a family of similar majestic madonnas, the most striking of which is in the Cathedral of Torcello near Venice (12th century). The tendency to depict icon-like, motionless mosaic figures isolated on a gold background has pre-Iconoclastic precedents, but from the 9th century onward it became a leading decorative principle.

Nineteenth-century drawings show that the decoration of Hagia Sophia also included comprehensive series of saints. Of these saints, which stood in rows on the nave walls above the galleries, only a few have survived. According to the drawings, those of the middle zone represented prophets, those of the lower, holy bishops. Higher up there

may have been a guard of angels and in the centre of the cupola, probably a mosaic of Christ. The disposition of the pictures, in other words, may have corresponded to that which at this time was being tried out especially for the new church architecture and which was to become the accepted system of decoration in the middle Byzantine churches.

Interplay of architecture and mosaic at Daphni

The monastery church at Daphni, near Athens, contains one of the best preserved decorations of this type (Figure 201). The building belongs to a category of central-plan

René Percheron—J.P. Zielo



Figure 201: Interplay between architecture and mosaics in the monastery church at Daphni, Greece, 11th century, crowned with a Byzantine dome mosaic Christ Pantokrator.

structures that had come into fashion and was to dominate for centuries both in Byzantium and in other areas under the influence of the Orthodox Church. The interior of the church at Daphni displays a layout which, compared with the wealth of detail of the early Christian period, appears single-minded and concentrated. In the centre of the dome is a medallion containing a colossal bust of Christ as Pantokrator, the All-Ruler. In the lowest part of the dome, separated from the medallion by a broad zone of gold, stand prophets with their scrolls. Further down, there may originally have been medallions with portraits of the Evangelists. In the four arches that carry the drum of the cupola are scenes from the life of Christ which, with eight more Christological scenes in the transepts, formed a cycle devoted to the central feasts of the church. The Virgin is represented in the apse, her guard of archangels on the side walls of the sanctuary. About thirty saints, depicted either as busts or as fulllength figures, fill the remaining wall space. In the vestibules are more scenes from the life of Christ and the remains of a cycle devoted to the life of the Virgin. Golden frames with floral ornaments surround the panels, and gold once covered every inch of wall between them.

The ensemble represents a visualization of the Christian cosmos, its effect created by an intricately conceived interplay of pictures and architecture. The worshipper who moves within this golden shell finds its world of pictures thoroughly involved with space. Space in fact fuses the decoration into one giant image, in which the ruler, hailed by the prophets surrounding him, presides in his sphere above the host of saints that people the lower part of the room.

Subtle spatial devices animate the individual pictures; figures of saints, their two-dimensionality emphasized by their outlines, appear in niches sunk in the wall or lean forward in the interior curves of arches. The 20th-century Austrian scholar Otto Demus, in studies on the aesthetics of middle Byzantine mosaic art, has coined the term space icons for this kind of imagery, in which the forms of architecture collaborate to make the solemnly stylized figures appear with unexpected tactility. As shown by Demus, the spatial element contributes to the narrative scenes also. In the four arches, for example, the hollow plane on which the scenes from the life of Christ unfold adds a dimension of spatial realism to the total image. This is most clearly to be observed in the Annunciation scene, where Mary and the Angel face each other across a stretch of real space. The figures share or are made to appear to share the room with the beholder.

The "classical system"

The "classical system," as this close interrelation of architecture and mosaic has been called, was probably perfected in the course of the 9th to 10th centuries, but the earliest fully preserved examples are from the 11th to 12th. Besides Daphni, Greece owns two more monuments of this kind, the monastery church of Hosios Loukas in Phocis and the Nea Moni on Chios (both 11th century). Similar churches are found in such widely distant places as Kiev (Hagia Sophia, 11th century) and Palermo (Martorana, *c.* 1150), both the products of strong Byzantine influence. The system, however, is not identical in any of these. The churches belong to the same general type, but their plans and elevations vary and thus require variations in this disposition of pictures as well.

The classical system with its emphasis on totality may have led to the gradual toning down of the many splendid effects of the earlier tradition for the sake of the equilibrium and clarity of the whole. At Daphni, for example, the rich, tapestry-like character of earlier mosaic has given way to a controlled, less sparkling range of tints. The reds and yellows are restricted, their function in the overall scheme taken over by the gold of the background. Sombre, often hard blues, greens, and violets are preferred to the lighter ones. Compared with the Hosios Loukas and the Nea Moni mosaics, which retain more of the older colour scheme (the latter almost to the point of brutality), the Daphni mosaics appear cool and intellectual, an impression further conveyed by their elegant style. Actually they belong to a new phase of Byzantine art which took its name from the dynasty of the Comnenus (AD 1081–1185). This style appears at its most refined in Hagia Sophia, Istanbul, in a panel depicting the Virgin flanked by the emperor John Comnenus II and his wife Irene. The practice of tilting the gold tesserae also seems to have been abandoned, for it is not found at Daphni nor in any of the mosaics that are examples of the fully developed classical system. Silver was reduced to the single role of depicting the light emanating from God and Christ. This drying out of the effects of light and colour was partly compensated for by a perfectionist setting and spacing of the tesserae.

Late Byzantine mosaics. The phenomenon called the Palaeologian Renaissance (from the dynasty of the Palaeologians, 1261–1453) led to a renewal of Byzantine mosaic art. The stylistic innovations that made themselves felt both in painting and mosaics of the late 13th and beginning 14th century bear witness to one of the most startling changes that ever took place within the framework of Byzantine culture. Bred by a vital humanism, which penetrated westward and laid the foundations for the Italian Renaissance, painting showed a predilection for perspective and three-dimensionalism. A peculiar vivacity invaded religious art, together with a sense of pathos and of the tragic. The results, as expressed in mosaics, were extraordinary.

To respond to the new trend, mosaicists recast their technique. The tessera size generally became smaller than it had been in earlier epochs; and contours lost their rigidity, became thinner, and were occasionally abolished. Colour was reintroduced in a manner that gives the Palaeologian works a striking likeness to the mosaics of the Early





Figure 202: (Left) Greek Festival Cycle miniature mosaic, 14th century. In the Museo dell'Opera del Duomo, Florence, Italy. (Right) Christ, detail from the Late Byzantine Deësis mosaic, south gallery, Hagia Sophia, Istanbul, *c.* 1300.
(Left) SCALA—Art Resource (right) Erich Lessing—Magnum

Christian period, which, one must suppose, in many cases served the artists as models. An interest in the optical effects of gold apparently returned but rarely, it seems, in the form of the tilting technique. On flat walls, the gold ground was sometimes set in a shell pattern, probably to enhance the play of light on the surface and to avoid a too-uniform brilliance. For domes, a densely ribbed form of cupola construction, which, when covered with mosaics, produces reflections of light that expand like rays from the central medallion toward the figures surrounding it, was preferred. Such domes are preserved in Kariye Cami, the former church of the Chora, at Istanbul, which was reconstructed and decorated as an act of piety by the logothete, or controller, Theodore Metochites in the second decade of the 14th century. Another superb example is found in Fetiye Cami (Church of the Virgin Pammakaristos) in the same city.

The feeling for colour, which is at its most refined in fragments from the decoration of the Church of the Holy Apostles in Thessaloníki (*c.* 1315) and at its most intense in the partly well-preserved cycles in the Kariye Cami, informs one of the greatest mosaic works of art, the Deësis panel in the south gallery of Hagia Sophia in Istanbul (Figure 202, right). In this same panel, the tilting technique reappears (in the cross arms of Christ's halo)—another indication of the retrospection inherent in late Byzantine art.

No mosaic in the true Palaeologian style has survived outside Byzantium. Reflections of it are found, however, in some of the 13th- and 14th-century works at Venice and in the mosaics executed by Pietro Cavallini in the apse of Sta. Maria in Trastevere in Rome (*c.* 1290–1300). Some of the characteristics of the style may have been brought to the attention of the Italian artists through portable mosaics, which despite their small size (generally about two by four to eight by ten inches [five by ten to 20 by 25 centimetres]) are imbued with many of the coloristic and technical features typical of monumental mosaics. Byzantine mosaic icons, the production of which was stimulated during the early Palaeologian era, were manufactured for personal devotion more than for the embellishment of churches and were exported in considerable numbers to the West or found their way there as gifts or booty in the politically troubled 14th and 15th centuries. In works whose quality can be compared with the most splendid of the Hellenistic *emblēmata*, extremely small tesserae, some measuring less than 0.04 inch (one millimetre) square, were assembled in wax or mastic on a board of fine wood. The tesserae material is often exquisite: silver, gold, and lapis lazuli and other semiprecious stones. The icons depict single figures such as saints, Christ, or the Virgin; single Christian scenes such as the Annunciation (Victoria and Albert Museum, London) and the Crucifixion (Staatliche Museen zu Berlin); or even the full Greek Festival Cycle (Figure 202, left).

Portable mosaics

MEDIEVAL MOSAICS IN WESTERN EUROPE

The prestige, both cultural and political, enjoyed by Byzantium in the Middles Ages led to a widespread imitation of its arts. Art objects in great number were imported to the West from Constantinople and other Greek centres. Individuals or communities outside the realm of Byzantium, however, were able to secure Byzantine artisans for the execution of monumental mosaics. Abbot Desiderius of the abbey of Montecassino in Italy, for example, called specialists in many crafts from Constantinople to decorate his new basilica (dedicated AD 1071). Among these were mosaic workers. Of particular importance is the fact that he took care to see that young local artists were trained by the foreigners. This was the pattern that was followed where Byzantine experts were temporarily called in.

Sicilo-Byzantine mosaics

The Norman rulers of Sicily, who vied with the Byzantines for control of the Mediterranean, molded their representational arts largely on those of the great Eastern power. The existence, at Palermo, of a central-plan church (Martorana) embellished according to the classical system has already been noted. In other 12th-century churches in Sicily, the Byzantine element is blended with western Mediterranean traits. Cappella Palatina, the palace chapel of the royal residence at Palermo (*c.* 1143 and later), for example, is a synthesis of a centralized middle Byzantine church and a basilica. The building therefore called for a hybrid program. According to Western custom, the mosaics of the basilical parts depict narrative cycles: scenes

from the Old Testament and from the lives of SS. Peter and Paul. In the centralized part of the church most of the features belonging to the classical system are at hand. There is a bust of the Pantokrator in the dome, surrounded by angels, but as a concession to the longitudinal disposition of the church, the Pantokrator reappears in the apse.

Martorana and Cappella Palatina were decorated by Byzantine artists, a fact borne out by the brilliant technique, the purity of the Comnenian style, and the adherence to Eastern iconographical prototypes. Two large basilicas are among the highpoints of this Sicilo-Byzantine flowering: the cathedral of Cefalù (*c.* 1148) and the church of Monreale, the last of the mosaic churches of Sicily (*c.* 1180–90). Demus has pointed out the extraordinary homogeneity of style and technique in the Monreale mosaics, which constitute the largest decoration of this kind in Italy. He has also shown that the Monreale mosaics are not executed in the refined and softly curved style that dominates in Cappella Palatina and at Cefalù. Monreale is infused with a more agitated and expressive style which, however, has nothing local or provincial about it. It was the late Comnenian style of Constantinople which had then reached Sicily—a testimony to the unbroken artistic contact that existed at this time between the Norman court and Byzantium.

Italo-Byzantine mosaics in Venice

Venice, for a long time commercially active in the eastern Mediterranean, enjoyed similar but more long-lasting artistic connections with Constantinople. A church that in the 11th century must have looked exotic as well as old-fashioned, St. Mark's was copied after the venerable Church of the Holy Apostles at Constantinople, an early Byzantine, many-domed type that had long since gone out of favour.

The mosaic work undertaken at St. Mark's lasted more than two centuries, from the end of the 11th to the middle of the 14th century. The many and marked stylistic disparities can be ascribed in part to the changes that affected Byzantine art during this period; but they may even have been caused by the freer, Western organization of the Venetian workshop, which allowed artists to develop and cultivate their own personal styles. Among the variety of styles, the Byzantine element is dominant, but it is modified by local tendencies and particularly by strong Romanesque impulses.

The workshops established in Sicily and Venice were active in the neighbouring areas. Mosaics in the cathedral at Salerno (*c.* 1190) and in the monastery at Grottaferrata near Rome (*c.* 1200) are regarded as products of the Sicilo-Byzantine school. The apse decoration of the cathedral at Ravenna (early 12th century), of which fragments survive, seems to have been the work of mosaicists from Venice and, in Florence, Venetian artists decorated the dome of the Baptistery (1225–1330). The much later mosaics on the facade of the cathedral at Prague (14th century) are also Venetian.

The Roman school

In the early Middle Ages, Rome had been able to maintain and defend a mosaic tradition of its own despite the Byzantine hegemony in the arts. At a period when Iconoclasm had loosened the ties between Byzantium and the West, at the end of the 8th century and the beginning of the 9th, the influence of the Roman school extended even into the Carolingian Empire. The apse mosaic of Germigny-des-Prés, France (between 799 and 818), is a product of this influence. In Rome, the pontificate of Paschal I (817–824) left three monumental decorations which constitute the best sources concerning the artistic intentions of this time: the sanctuary mosaics of the churches of Sta. Cecilia in Trastevere, Sta. Maria in Domnica, and Sta. Prassede (attached to the latter is the domed chapel of S. Zeno, also fully decorated with mosaics). The iconographic schemes of these 9th-century churches largely reflect local 5th- and 6th-century church decorations. Also remarkable is the return of many of the technical idiosyncrasies of the early Christian period, including the employment of large tesserae for faces and dress and the use of glass for all parts of the composition. The setting of the tesserae is, however, loose and disorganized; and their varying shapes suggest that they are reused cubes taken from older monuments. Stylistically, the mosaics of the Paschal period, with their extreme two-dimensionality, are related to Byzantine mosaics of the 9th century (the cupola of Hagia Sophia, Thessaloníki), but technically the differences could not be greater.

Little is known of the art in 10th and 11th century Rome. But from the 12th century, a group of works testify to continued inbreeding. The apse decoration of S. Clemente (the 1st half of the 12th century), for example, contains a scroll pattern (rinceaux) reproduced from a 4th-century decoration in a technique that in many respects resembles that of the Paschal mosaics. This isolation, or resistance to foreign influence, seems to have been broken in the 13th century. In 1218, Pope Honorius III asked the doge to send craftsmen for the decoration of S. Paolo Fuori le Mura (1218). Several important mosaics from the later part of the same century reflect the trends current at that time in Byzantine and Italo-Byzantine mosaics. The mosaics by Jacopo Torriti in the apse of the basilica of Sta. Maria Maggiore (*c.* 1290–1305) are among the finest of these. They show a mingling of Western medieval and Early Christian iconographical features, such as a scene of the crowning of the Virgin surrounded by the scrolls of a fleshy, classicizing rinceaux; but the lower zone with scenes pertaining to Mary reflects Byzantine influence, an influence also seen in the technique and colour scheme of the mosaics.

Floor mosaics had a renaissance in the West that was unmatched in the Eastern Empire. In the early Middle Ages, the Byzantines developed their particular form of floor covering consisting of a geometrical mosaic made up of pieces of marble of various sizes and shapes (*opus sectile*) usually with some tessera work added for special coloristic touches. This art, called Cosmati work, spread to the West, in which, however, there was also a revival of the tesselated pavement. This pavement, which had survived the Dark Ages in a very primitive form, reemerged in Italy in the 11th century in greater splendour. There are impressive remains of such floors in many of the larger Italian churches. The fashion spread to other parts of Italy and even to France and Germany. In France, where large floors were produced in quantity in the 12th century, fragments of outstanding quality are found; for example, in Saint-Nicolas' at Reims (last half of 12th century). Cologne seems to have been a leading centre for this art in Germany. The style of the floors is usually one of simple outlines and light colours, though in some cases figures and ornaments appear against dark ground. The programs draw their inspiration from many sources, such as textiles, early floor mosaics, and the sculptural ornamentation of churches. An exceptionally well-preserved example is found in Otranto in the Italian province of Apulia, now Puglia (1163–66), where vast floors depict scenes from the Old Testament and mythology (the ascension of Alexander), representations of the Zodiac, and of the labours of the month. A profusion of monsters and fantastic animals fills out the picture of a decoration cast in the Romanesque iconographic tradition.

Cosmati work

RENAISSANCE TO MODERN MOSAICS

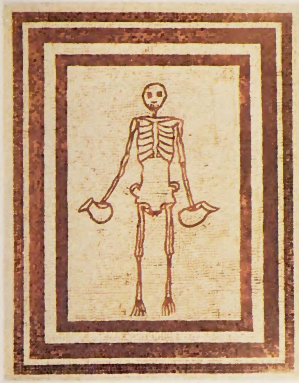
With the downfall of Byzantium in the 15th century, there perished that milieu in which mosaic had been constantly cultivated and had undergone continuous renewal in response to changing patterns of religious and cultural life. The art lost another foothold in Italy at the beginning of the same century, when changing attitudes about the world and about the function of art eliminated the very bases upon which mosaic had been built. One of the conventions against which the artists of the Renaissance, who were striving for pictorial realism, most strongly rebelled was the use of gold, the other-worldly element most typical of mosaic art.

The decline of mosaics

Although mosaic continued to be used to a certain extent as church decoration, it was a changed art. Some of its traditional glitter was retained, but essentially mosaics became imitations of painting. These imitative intentions were disastrous and led to the loss of knowledge of how to blend colours and handle materials. In earlier mosaics, there undoubtedly had been a distinction between the leading artist of the project, who drew the composition



Dionysus on a Tiger, from the House of the Faun, Pompeii, 2nd century BC. In the Museo Archeologico Nazionale, Naples.



Skeleton of a Cup-Bearer, from the House of the Faun, Pompeii, 2nd century BC. In the Museo Archeologico Nazionale, Naples.



Detail from the mosaic of the Nile, known as the "Barberini" mosaic, from the Sanctuary of Fortuna Primigenia, Palestrina, c. 80 BC. In the Museo Archeologico Nazionale, Palestrina, Italy.

Roman mosaics



Interior court with mosaic of Neptune and Amphitrite, from the House of Neptune and Amphitrite, Herculaneum, 1st century AD.

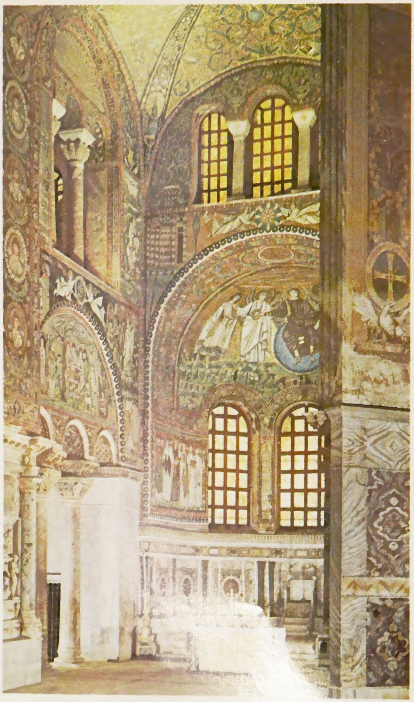
Plate 11: (Top left, bottom left) P. Marzari, (top right, bottom right) SCALA—Art Resource

Early Christian, early Byzantine, and Islamic mosaics



The Good Shepherd, Mausoleum of Galla Placidia, Ravenna, *c.* 450.

Mosaics in the church of S. Vitale, Ravenna, *c.* 546–547.





Floor mosaic from the bath of the Palace of Khirbat al-Mafjar, Jericho, *c.* 743.



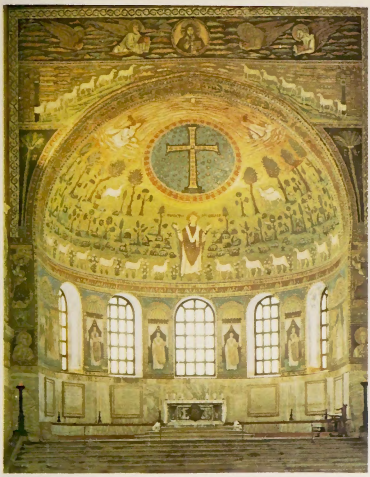
Detail from the mosaics of the great mosque of Damascus, Syria, 715.

Plate 12: (Top) SCALA—Art Resource; (left) Shostal—EB Inc.; (centre right) Micha Bar-Am, (bottom right) Cipclone—Rapho/Photo Researchers

Dome of the Baptistery of the Orthodox, Ravenna, c. 450.



Apse in the church of S. Apollinare in Classe, Ravenna, second half of the 6th century.





Crossing of the Red Sea, in the church of Sta. Maria Maggiore, Rome, first half of the 5th century.



Detail from the story of Jonah, pavement mosaic in the cathedral at Aquileia, second decade of the 4th century.



Panel depicting the Virgin and Child with the emperor John Comnenus II and the empress Irene, in Hagia Sophia, Istanbul, *c.* 1118.



Detail of the floor mosaics in the Great Palace of the Emperors, Istanbul, *c.* 600.

Middle and late Byzantine and medieval mosaics

Crowning of the Virgin, detail of the apse mosaic by Jacopo Torriti in Sta. Maria Maggiore, Rome, 1295.





Apse mosaic of Germigny-des-Pres, near Orleans, France, 9th century.

Plate 14: (Top) Andre Held—J.P. Ziolo, (centre left) Ara Guller, (bottom left) Giraudon—Art Resource, (bottom right) SCALA—Art Resource
Plate 15: Shostal—EB Inc.



Interior mosaics of St. Mark's, Venice, 11th through 13th centuries.



Modern mosaics

Dome of St. Peter's, Rome, mosaics after cartoons by Cavalier d'Arpino, begun in 1576.



Sculpture mosaics by Jeanne Reynal, 1970–71.



Vault of the Chigi Chapel, by Luigi da Pace after cartoons by Raphael, 1513. In the church of Sta. Maria del Popolo, Rome.

Mosaics at University City, Mexico City: (left) Central Administration Building, mosaic by David Alfaro Siqueiros, 1952–53; (right) Library, mosaic by Juan O'Gorman, 1951–53.

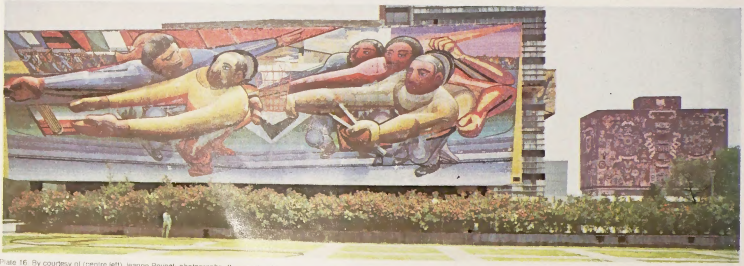


Plate 16. By courtesy of (centre left) Jeanne Reynal; photographs, (top, centre right) SCALA—Art Resource, (bottom) Shostal—EB Inc.

and oversaw the execution, and the ordinary setters of the tesserae. The leading artist, however, almost certainly took a hand in the setting of special parts and was thoroughly trained in the technical side of the production. Now the preparatory work was divorced from the execution: the artist submitted his cartoon and left its transposition into mosaic to artisans. This drew the lifeblood from the art and caused its degradation.

In Italy, many of the great painters of the 15th and 16th centuries delivered designs for decorations in mosaic. Best known among these decorations are the works of the Venetian Luigi da Pace after Raphael's cartoon, in the dome of the Chigi Chapel in Sta. Maria del Popolo in Rome (1516), and the mosaics made after the cartoons of Titian, Tintoretto, Giuseppe Salviati, and Paolo Veronese to complete the decoration of St. Mark's in Venice. Among the greatest single undertakings of this kind was the decoration of the dome of St. Peter's in Rome, executed in the last quarter of the 16th century from the cartoons of Cavalier d'Arpino. St. Peter's also displays some of the most technically striking mosaic reproductions of paintings ever executed—the much admired altar pictures after originals by 16th- and 17th-century masters. Created for the completion and care of the large mosaics of the two great churches, the workshops attached to St. Peter's and to St. Mark's gradually became centres for the manufacture of mosaics. From them, artists were summoned for decorative work in all parts of Europe. The school of mosaics in the Vatican and the workshops in Venice still have a considerable share in the field, together with the school more recently set up for the restoration of the mosaics at Ravenna.

Nineteenth-century historicism and the breaking down of Neoclassicism's contempt for Byzantine art led to an increasing interest in and demand for mosaics. Improvements in the technique of prefabrication according to the indirect method and in the manufacture of tessera material led to a veritable mass production, which has put its mark on countless churches, town halls and opera houses. Shrill colours and a gleaming, metal-like surface characterize many of these works.

The modern revival of mosaics had several causes. Scholarship and tourism made the monuments of ancient and medieval mosaics available to an art-loving public. Painting, since the last third of the 19th century engaged in the exploitation of colour, at the turn of the century focussed on the problems of colour as the expression of psychological qualities rather than of the external world. Expressionism, which opened the eyes of artists to the art and artifacts of foreign and distant cultures, also turned their interest towards medieval mosaics. The abstract element which these mosaics contain and which springs from the latent conflict between the design and the tessera pattern made them particularly attractive to artists of the earliest decades of this century such as Marc Chagall and Giovanni Serverini. The texture of mosaic was also an attraction. An American mosaicist, Jeanne Reynal, for example, created abstract compositions in which texture is emphasized by a combination of granulated, pebble-sized, and normal tesserae, sparsely spread over a coloured base of portland cement. Many of these mosaics are small and are hung on the wall like paintings.

Mosaic's smooth yet faceted surface is ideal for decorating the large, unbroken surfaces of modern architecture. The greatest modern use of mosaic as architectural decoration is in Mexico, a country with a long tradition of folk mural painting. Realizing the potential of the medium for public enjoyment and education, the government in the 1930s and 1940s commissioned many murals with historical and political themes for public buildings. Later, it became desirable to decorate the exterior walls of buildings, and mosaic was the logical alternative to the less durable murals. Often mosaics were designed by mural painters such as Diego Rivera who, in 1953, designed the immense mosaic on the facade of the Teatro de los Insurgentes. Francisco Eppens also used historical themes in his mosaic decorations of the schools of medicine and dentistry at the National Autonomous University of Mexico (1957), as did Xavier Guerro in the Cine Ermita in Mexico City. Carlos Mérida, however, created abstract mosaic designs in the Reaseguras Alianza in Mexico City. Among the most prolific Mexican mosaicists was the architect-muralist Juan O'Gorman. Of his many mosaic works, the most important is on the exterior walls of the library of the National Autonomous University of Mexico (1951–53), which exemplifies the monumentality of which mosaics are capable. Other works executed by O'Gorman include mosaics on the SCOP, or Secretaría de Communicaciones y Obras Públicas (1952), and a stone mosaic on the facade of the Posada de la Misión Hotel in Taxco. In 1950, O'Gorman began to decorate his own house in Mexico City with phantasmagoric images and symbols from Aztec mythology.

Renaissance of the mosaic art in Mexico

Mosaic in the strict sense of the word is an art in transition. The conventional tessera mosaic is still largely in use, mainly because of the efficient production methods of modern mosaic firms; yet the distinction between this kind of mosaic and other, mosaic-like techniques is slowly being dissolved. A mixture of cubes and larger pieces of glass, ceramics, or stone is one of the variants that occurs, cubes and *objets trouvés* ("found objects") another. After World War II, experiments seem to have moved in the direction of an employment of larger, less regularly cut units. Today, mosaic is just one out of a very wide range of techniques that have as common principle the piecing together of a surface covering with the use of different durable or nondurable materials. (P.J.N.)

PRE-COLUMBIAN MOSAICS

Maya mosaics

The art of mosaic in pre-Columbian Central America was marked by a combination of great technical skill and widespread use. The representation of a mosaic mask on a stela at Seibal (an upright, freestanding stone slab functioning as a commemorative monument), Guatemala (AD 590), established the early use of the technique in Maya territory, but it became best known from the few specimens surviving from the time of the Aztec empire (*c.* 1376–1519) and from descriptions of others left by the Spanish conquerors. The Mexican lapidaries worked with obsidian, garnet, quartz, beryl, malachite, jadeite, marcasite, gold, mother-of-pearl, and shell, but turquoise above all constituted their favourite material; the excessive richness of the religious ceremonial gave wide range to its employment in ritual paraphernalia of all sorts. The incrustation was laid upon wood, stone, gold, shell, pottery, and possibly leather and native paper and was

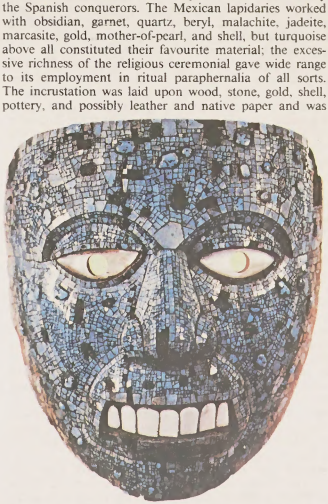


Figure 203: Mask of Quetzalcoatl, turquoise tesserae on wood with mother-of-pearl, Mixtec, from Mexico, 14th–15th century. In the British Museum.

By courtesy of the trustees of the British Museum

held in place by a tenacious vegetal pitch or gum or a kind of cement.

Masks, shields, helmets, knife handles, staffs, collars, medallions, ear plugs, leggings, mirrors, animal figures, and cult statues received a covering, in whole or part, of small and irregularly shaped pieces of highly polished turquoise, cut to fit tightly together so as to form a brilliant green surface, varied at times with cabochons (a gem cut in convex form, highly polished but not faceted) of turquoise or other material. Most striking and best preserved of the surviving two dozen major specimens of this art are a mask in the British Museum, London (Figure 203), and a shield in the Museum of the America Indian, Heye Foundation, New York City. Minute pieces of turquoise, studded with cabochon turquoises, completely cover the cedar mask and are laid in symmetrical lines around the eyes and mouth and on the nose; eyes and teeth are of shell inlay. Over the wood of the shield, one panelled and three circular borders of mosaic frame a scene that may relate to the worship of the planet Venus; it is estimated that nearly 14,000 pieces of turquoise make up the decoration. A Nahuatl story of a hall at the Toltec city of Tula, the walls of which were covered with fine mosaic, may belong to legend, but a monumental use of the technique was achieved in the mosaic-like treatment of the exterior wall casing of certain buildings. Those at Mitla in the state of Oaxaca are outstanding; bands and panels of simple but striking geometric ornamentation were produced by fitting together small stones of different shapes and sizes, tenoned back into the rubble mass of the wall. Each stone was cut for the spot it occupied, and some were more deeply imbedded than others so that the designs stand out in sharp relief. The effective simplicity of design and precision of workmanship at Mitla are not matched on the elaborate Maya facades—at Uxmal and Chichén-Itzá in Yucatán, for example—where, along with geometrical designs, animal forms also occur.

Mitla and Oaxaca

(F.O.Wa.)

STAINED GLASS

All coloured glass is, strictly speaking, "stained," or coloured by the addition of various metallic oxides while it is in a molten state; nevertheless, the term stained glass has come to refer primarily to the glass employed in making ornamental or pictorial windows. However, the singular colour harmonies of the stained-glass window are due less to any special glass-colouring technique per se than to the exploitation of certain properties of transmitted light and the light-adaptive behaviour of human vision. Rarely equalled and never surpassed, the great stained-glass windows of the 12th and early 13th centuries actually predate significant technical advances in the glassmaker's craft by more than half a century. And much as these advances undoubtedly contributed to the delicacy and refinement of the stained glass of the later Middle Ages, not only were they unable to arrest the decline of the art, they may rather have hastened it to the extent that they tempted the stained-glass artist to vie with the fresco and easel painter in the naturalistic rendition of his subjects.

Neither painting on stained glass nor its assembly with grooved strips of leading is an indispensable feature of the art. Indeed, the leaded window may well have been preceded by windows employing wooden or other forms of assembly such as the cement tracery that has long been traditional in Islāmic architecture; and the single most important technical innovation in 20th-century stained glass, slab glass and concrete, is a variation on the earlier masonry technique.

Elements and principles of design

Of all the painter's arts, stained glass is probably the most intractable. It is bound not only by the many light-modulating factors that affect its appearance but also by comparatively cumbersome, purely structural demands. And yet no other art seems so little earthbound, so alive, so intrinsically beguiling in its effect. This is because stained glass, far more directly and intensively than other media, exploits the interaction between two highly dynamic phenomena, the one physical and the other organic. The physical factor is light and all of the myriad changes in the general light level and the location and intensity of particular light sources that occur as a matter of course not only from moment to moment but from place to place—a prairie to a forest, a greenhouse to a dungeon. The other phenomenon is the spontaneous light-adaptive process of vision, which seeks to maintain orientation in all luminous environments.

Architecture, by determining the apparent brightness value of the light seen through its window openings, always establishes a definite scale of brightness values with which the stained-glass artist must work. Because the light that penetrated the interior of the 12th- and early 13th-century church took on a brilliance, even harshness, in contrast to the surrounding darkness, the artisans of the period logically composed their windows with a palette of deep, rich colours. When for doctrinal or economic reasons only clear glass could be used, it was decorated with a fine opaque mesh of grisaille, or monochromatically painted ornament, that effectively broke up and softened the light. Later, as the walls of the churches were opened up to admit more and more light, the difference between the interior and exterior light levels was no longer great enough to illuminate the dense, saturated rubies and blues of the earlier period. In the 14th and 15th centuries, generally higher keyed, drier, and more muted colour harmonies were developed. This reflected a growing preference for lighter, less awesome effects and an actual limitation that the architecture of the time imposed upon the medium of stained glass.

Role of light

The static elements of the glass and its architectural setting are modified by the element of change inherent in natural light. A seemingly endless spectrum of changes in the appearance of stained glass is a result of the changes in the intensity, disposition, atmospheric diffusion, and colour of natural daylight. The luminous life of stained glass, therefore, can best be observed by watching the organic effect of light on the window through the course of a day. If one were to enter the Cathedral of Chartres just after sunrise on the morning of a clear day, it would be to the east windows, especially those in the clerestory, that his eyes would first be drawn. They alone will have come fully to life and all of the others will still seem to half-exist in a kind of hushed twilight. Gradually, as the sun rises in the sky, these windows will become more luminous. Then the east windows will begin to lose their earlier brilliance to those all along the south flank of the cathedral, which by midday will be fairly aglow from the direct rays of the sun. The light streaming through the south windows, however, will have raised the light level inside the north windows opposite them sufficiently to create a distinct, though by no means unpleasant, muting of the radiance of the latter. If the sun at this point disappears behind a cloud and the sky becomes generally overcast, the appearance of all of the windows is immediately and dramatically altered. Because the light, now diffused, comes more or less equally from all directions, the south windows will lose some of their earlier brilliance and vivacity and the north windows will recover theirs. The overall atmosphere of the cathedral is distinctly cooler and graver in its effect, and more than ever before one begins to become aware of absolute differences in the tonality of the various windows themselves. The grisaille windows in the east end of the cathedral, the highly keyed 15th-century window in the Vendôme Chapel in the south aisle of the nave, and the three 12th-century windows over the great west portal all stand out as being substantially more luminous than the rest. If, late in the afternoon, the sun reappears, the viewer is treated to an extraordinary spectacle as the blues in the west windows, by far the most intense in the cathedral,

are further emblazoned by the direct rays of the sun. Should the main doors of the cathedral be opened, the direct rays of the late afternoon sun, streaming halfway down the nave of the cathedral, will cast a blinding pall over all the windows within their vicinity until the doors are closed once more. Then as the sky begins to redden with the setting sun, the intense 12th-century blues in the west windows lose their former intensity, and the warmer colours, especially the rubies, become so fiery and assertive that they seem almost to have displaced the blues as the predominant colour in the windows. Finally, when the sun is gone the whole cathedral is plunged once more into a deep twilight, which gradually diminishes until there is no light at all.

Insofar as stained glass may be considered an art of painting, it must be considered an art of painting with light. Whatever techniques or materials it may employ, its own most unique and indispensable effects are always the product of colouring, refracting, obscuring, and fragmenting light.

Materials and techniques

Contrary to popular belief, the glassmaker and the stained-glass artist could seldom have been the same person even in the earliest times; in fact, the two arts were rarely practiced at the same location. The glassmaking works was most readily set up at the edge of a forest, where the tremendous quantities of firewood, ash, and sand that were necessary for the making of glass could be found, whereas the stained-glass-window-making studios were normally set up near the major building sites. The stained-glass artist, thus, has always been dependent upon the glass-maker for his primary material. Coloured with metallic oxides while in a molten state—copper for ruby, cobalt for blue, manganese for purple, antimony for yellow, iron for green—sheets of medieval glass were produced by blowing a bubble of glass, manipulating it into a tubular shape, cutting away the ends to form a cylinder, slitting the cylinder lengthwise down one side, and flattening it into a sheet while the glass was still red hot and in a pliable state. It was then allowed to cool very slowly in a kiln so that it would be properly annealed and not too difficult to cut up into whatever shapes might be required for the design. Since these sheets of glass, with the exception of a type known as flashed glass, were intrinsically coloured with one basic colour throughout, changes from one colour to another in the design of a window could be effected only by introducing separate pieces of glass in each of the requisite colours.

Role of the glassmaker

Whether by accident or by deliberate intent, the glass made in the 12th and 13th centuries had almost the ideal combination of crudity and refinement for stained glass. The sheets, 10 by 12 inches (25 by 30 centimetres) in size, were both flat enough and thin enough to be cut very accurately into the necessary shapes, yet still variable enough in thickness (from less than $\frac{1}{8}$ inch [three millimetres] to as much as $\frac{5}{16}$ inch [eight millimetres]) to have rich transitions in the depth of their colours. With the progress of glass technology in the Middle Ages and Renaissance came the ability to produce larger, thinner, and flatter sheets of glass in a considerably larger range of colours than had been possible in the 13th century. At each distinguishable stage in this development, however, the glass became less visually interesting as an aesthetic element in its own right. The Gothic Revivalists later recognized this effect, and in the mid-19th century they initiated a return to the earlier methods of producing glass. They developed the so-called "antique" glass, which is remarkably similar in colour, texture, and shading to the glass that was used in the 12th- and 13th-century windows. "Antique" glass remains the basic material used in stained-glass windows to this day.

Traditional techniques. The art of stained glass is the translucent offspring of such earlier art forms as mosaic and enamelling. From the mosaicist came the conception of composing monumental images out of many separate pieces of coloured glass. Cloisonné enamelling probably inspired not only the technique of binding these pieces together with metal strips but that for treating the strips themselves as a positive design element. From the enamellers must also have come the near-black vitreous enamel made from rust powder and ground glass that was mixed with a mild water-based glue to form a paint. This could be used to render more or less opaquely onto glass the details of figures, ornaments, and inscriptions.

The technique of making stained-glass windows is first described in the *Schedula diversarum artium*, a compendium of craft information probably written between 1110 and 1140 by the monk Theophilus (tentatively identified as the 12th-century goldsmith Rugerus of Helmarshausen). First, a full-sized cartoon, or line drawing, of the window was painted directly onto the top of a whitewashed table, showing the division of the various colour areas into individual pieces of glass (Figure 204). Next, sheets of glass

Brunel—Zoio

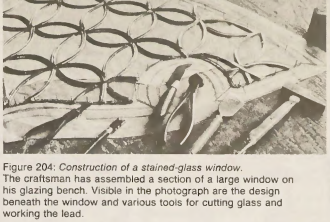


Figure 204: *Construction of a stained-glass window.*
The craftsman has assembled a section of a large window on his glazing bench. Visible in the photograph are the design beneath the window and various tools for cutting glass and working the lead.

of the appropriate colours were selected and from these pieces were cut, or, more accurately, cracked away with a red hot iron. By applying the hot iron to the edge of the sheet it was possible to start a crack that could then be guided more or less in the direction in which the iron was moved, thus enabling the glazier to break away from the sheet of glass a piece of approximately the right shape and size. This he would then further shape by "grozing," or crumbling away bits of glass from its edges with a notched tool known as a grozing iron. When all of the pieces were thus accurately cut to shape, with due allowance between pieces for the leads that would join them together, the details of the design were painted onto the glass wherever necessary with vitreous enamel. The pieces were then placed in a kiln and fired at a temperature just hot enough to fuse the enamel to the glass. This done, the windows were ready for assembly with grooved strips of lead that look in cross section like the letter H. The glazier would begin by butting together on his workbench two long strips of lead, to form a corner of the panel. He would then set the corner piece of glass in place between these two leads and cut another strip of lead just long enough to surround the rest of the piece. Against this lead he would then be able to set the next piece of glass, and so on across the panel, until it was completely assembled on the glazing bench. The joints between the leads were then soldered, the panel was waterproofed by rubbing a putty compound under the leads, and it was ready for installation.

Cutting and leading techniques

Because of the flexibility of the leading it was found necessary to divide all but the very smallest windows into a series of separate leaded panels and to insert iron framing members, or armatures, between the panels. In the earliest single-figure lancet windows, such as the "Prophets" in Augsburg Cathedral (Figure 205, left), the divisions tend to be purely functional. Very soon, however, more ambitious windows became much too large to be handled in this manner. Whereas the Augsburg "Prophets" measure only about 12 square feet (1.1 square metres) in area, the Poitiers Cathedral "Crucifixion" window (Figure 205, centre) contains approximately 175 square feet (16.3 square metres) of stained glass, and the "Life of Christ" in Chartres contains more than 250 square feet (23.2 square metres). A much more elaborate system of subdivisions in the window opening, consisting of vertical as well as

horizontal members, was developed. These systems of supports often formed a geometric pattern that was incorporated in the overall design of the window. In fact, it was the ingenious conversion of this structural necessity into a positive design element that set the stage for the creation of the medallion windows of the great Gothic cathedrals. By utilizing these armatures to delineate the principal ornamental subdivisions of the windows, as in the Chartres "Good Samaritan" (Figure 205, right), the glass painters were able to fuse a complex didactic imagery and an austere architecture into one of the most compelling artistic unities of Western art. At the same time, particularly in the upper levels of a church, stone mullions began to be employed for the same purpose. The most spectacular examples are the great rose windows, in which masonry is so literally dissolved into fenestration, and the individual window opening so completely absorbed into the overall pattern, as to defy any meaningful distinction between window and wall. This perfect fusion of image, ornament, and structure, with each deriving strengths from the others that none would ever have alone, was one of the most significant turning points in the history of stained glass. From this point on the relation between stained glass and architecture begins to decline. The aims, techniques, and achievements of the stained-glass artist begin to resemble those of the fresco and easel painters, and it is by the standards applicable to the latter that the stained glass of the 14th, 15th, and 16th centuries must be judged.

Developments in the 14th century. The first significant developments in the glass painter's craft appear to have been made more or less simultaneously in the early years of the 14th century. Glass in a range of previously unavailable secondary colours—smoky ambers, moss greens, and violet—becomes generally available for the first time. The technique of staining glass yellow by painting it with silver salts is discovered (Figure 206). The glass painters also begin to develop a number of techniques for shading or modelling forms with vitreous enamel by applying translucent matts of halftone to the whole surface of the window and delicately brushing it away where highlights are desired. Darker shading is sometimes reinforced by painting on the outer as well as the inner surface of the glass. The uses of line also become increasingly refined and versatile, especially in the 15th century.

Painting and grinding techniques

To these refinements of the craft was added one wholly new technique, the abrasion of flashed glass. Ruby glass, whose unique composition made this technique possible, was a laminated glass, although it appears to be coloured

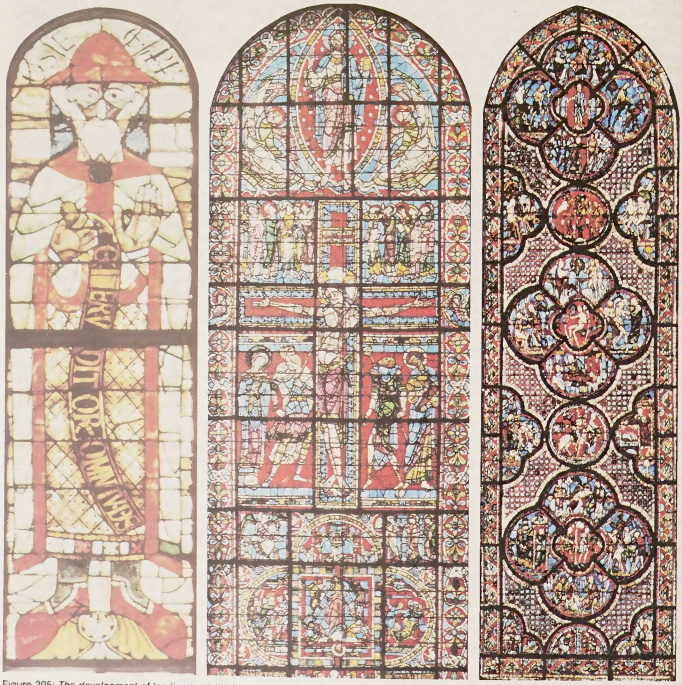


Figure 205: *The development of leading in stained-glass windows.*
(Left) "The Prophet Hosea," single-figure window, *c.* 1125. In Augsburg Cathedral, Germany.
(Centre) "The Crucifixion," elaborately divided window, *c.* 1165. In Poitiers Cathedral, France.
(Right) Scenes from the life of the "Good Samaritan," medallion windows, first quarter of the 13th century. In the south aisle of the nave of Chartres Cathedral, France.
(Left) Edana-Verlag, (centre) Lauros—Giraudon from Art Resource (right) Giraudon—Art Resource

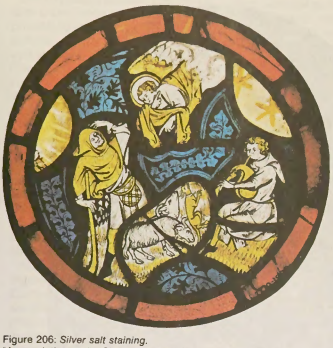


Figure 206: *Silver salt staining.*
"Annunciation to the Shepherds." English 14th-century stained-glass window in which silver salts have been used to stain the glass shades of yellow, and the reds are "streaky ruby" glass. In the Victoria and Albert Museum, London.
By courtesy of the Victoria and Albert Museum, London; photograph, J.R. Freeman & Co. Ltd

intrinsically throughout like all of the other glass in the early windows. Because the metallic agent used to produce its colour was so dense, all but the thinnest films of ruby were opaque. To obtain sufficient translucency, either the glassmaker had to suspend striations of ruby in a clear glass, thereby creating the "streaky rubies" of the early 13th century, or the glass was "flashed"; that is, clear glass while still pliant was dipped into molten coloured glass, thus coating its surface with a thin film of colour. Detailed effects, unhindered by intricate leading, could then be achieved by grinding away portions of this coloured film, first on ruby glass and then on other colours deliberately "flashed" for this purpose. To these colours could now also be added the silver salts stain in tones of yellow ranging from the palest canary tint to a deep fiery amber, depending on how heavily the stain was applied and how thoroughly it was fired. The whole gamut of more or less translucent tonalities that could be created with vitreous enamel were also used. Taken altogether, these techniques when used in combination represented a considerable liberation of stained glass from what was increasingly considered to be the "tyranny" of the lead line.

The technique of grinding flashed glass was first practiced in the late 13th and early 14th centuries; one of the earliest extant examples is in the church at Mussy-sur-Seine in France, where the windows have a blue groundwork covered all over, or diapered, with ruby roses with white centres, each rose being a single piece of glass. This type of work, however, was not common until the 15th and 16th centuries.

Later developments. At the end of the 15th century a whole new range of vitreous enamels was developed, and by the middle of the 16th century the technique of painting in enamel colours on glass began to be of major importance. In this method, granulated coloured glass of the desired colour is mixed with a flux of clear ground glass and fired onto the surface of the glass. Enamel painting was not altogether successful either technically or aesthetically, since the colours thus created were translucent rather than transparent, generally pallid, and of uncertain durability. Political disturbances in the mid-17th century created a scarcity of coloured glass throughout Europe, and gradually the traditional use of coloured glass was replaced by the new technique.

Modern techniques

Between the 16th and 20th centuries the developments in the craft of making stained-glass windows were purely utilitarian. In the 16th century the diamond glass cutter was invented, and in the 18th century hydrofluoric acid was introduced as a means of etching flashed glass. In the 19th and 20th centuries, gas and electric kilns and soldering irons were used, as were plate-glass easels upon which stained-glass panels could be temporarily mounted for painting before they were leaded. The largest palette of glass—the widest range of colours, textures, and thicknesses that the art has ever known—was also developed in the 20th century. Contemporary technical innovations include the slab glass and concrete windows developed in France in about 1930, where glass set in concrete provides an alternative to leading. In the mid-20th century such experimental techniques as bonding glass to glass with transparent resin glues were developed. Measured purely by technical standards, contemporary stained glass has never been rivalled in its versatility as an instrument of artistic expression.

Subject matter

In the Middle Ages ecclesiastical art was primarily didactic. The subjects painted in the windows played an important part in the expounding of the Scriptures and the glorification of the church and its saints.

Iconographic programs

The iconographic program of medieval stained-glass windows for ecclesiastical buildings is a product of several factors. To begin with, the cruciform plan of the churches themselves created four focal areas. Each area, by its architectural form and orientation to the sun, tended to elicit the development of certain subjects or types of subject. In Chartres, for example, the five central windows of the choir clerestory and the north rose window (Figure 207, right) are consecrated to the Virgin, the south rose window to the glorification of Christ, and the west rose window to the Last Judgment. In Bourges Cathedral the huge figures of the Apostles in the south clerestory are paired off against the prophets in the north clerestory, the representatives of the New Testament thereby receiving the full light of the sun and their Old Testament counterparts the more crepuscular light of the north sky. The great rose windows, whose circular form is itself cosmological in its implications, are invariably devoted to cosmological themes: the Last Judgment, the Apocalypse, the cycle of the year as expressed in the signs of the zodiac, the glorification of Christ and of the Virgin as the rulers of heaven. On the other hand, one of the reasons that the theme of the Jesse tree remained popular throughout the Middle Ages was that it lent itself to such a rich variety of ornamental treatments (Figure 207, left). And finally there was the will of the donors of the windows, whose personal preferences determined the subjects of many excellent works that clearly cannot be related to any comprehensive iconographic program. Some idea of the scope of these medieval enterprises can be indicated by the fact that Chartres, by no means the largest of the cathedrals, contains more than 27,000 square feet (2,500 square metres) of stained glass, in 176 windows. Of the 64 windows on the lower level, all but a few are medallion windows, which contain anywhere from 20 to 30 or more separate pictorial compositions; and the three rose windows, each more than 40 feet (12 metres) in diameter, are vast composite creations. The work of at least nine separate master designers has been distinguished in the windows of the cathedral, which was completely glazed in less than 40 years, between about 1203 and 1240.

It must be assumed that clerics supplied the master glazier with a program to which he had to conform. A 12th-century manuscript in the British Museum contains a series of circular drawings illustrating the life of St. Guthlac. These drawings might have been intended as a model for a glazier, but the scenes could equally well have been expressed in wall paintings, sculpture, or metalwork. There is more complete knowledge for the later Middle Ages. The glazier was given written instructions from which to prepare provisional sketches that were submitted for the patron's approval before being redrawn in actual size to form the final cartoon. The provisional sketch was known as a *vidimus* (literally, "we have seen"). One example of such written instructions is the program for a window given by Henry VII to the Grey Friars Church at Greenwich, England.

The *vidimus*



Figure 207: (Left) The Jesse tree window, by the Le Prince family, *c.* 1522–25. In the church of Saint-Etienne, Beauvais, France. (Right) The north rose window, dedicated to the Virgin, and five lancets, with Melchizedek, David, St. Anne holding the baby Mary, Solomon, and Aaron, *c.* 1230–35. In Chartres cathedral, France.
(Left) Jean Roubier—EB Inc. (right) Promophot—Zizlo

There is ample evidence to show that by the 14th century it was the practice of glaziers to have a stock of finished cartoons, executed on parchment or paper, which could be adapted for different glazing schemes. That these cartoons were used and reused over a long period can be deduced from the will of a York glazier, who died in 1450, in which he bequeathed to his son all his cartoons.

It is evident that in the later Middle Ages the master glazier's workshop was a highly organized enterprise, capable of producing various classes of designs, according to the expense his patrons were prepared to incur. Although the donor, cleric or layman, exercised considerable influence over the choice of subject and its manner of representation, the finished design was essentially the creation of the master glazier. The latter was often an artist in his own right, expressing in the formal language of his own technique the artistic aspirations of his time. (R.So.)

Periods and centres of activity

The evolution of the stained-glass window was a slow process. Both texts and excavation testify to the existence of stained-glass windows before the 12th century, but the textual references are too brief and nontechnical to give any clear picture of how the art evolved. The writings of the Fathers of the Latin Church—Lactantius (*c.* AD 240–*c.* 320), Prudentius (AD 348–after 405), and St. Jerome (before 420)—mention coloured glass windows in the early Christian basilicas. The 5th-century poet Sidonius Apollinaris described glazed windows in Lyon, France. Pope Leo III (795–816) is recorded to have provided windows of different coloured glass for St. Paul's basilica at Rome. Glazed church windows were widespread in pre-Carolingian Europe in the the wealthiest establishments: the Cathedral of York in England was glazed as early as 669. On the site of the Abbey of Monkwearmouth in Sunderland, England, a number of pieces of window glass dating from the late 7th century were found. Coloured green, blue, amber, and red, the edges of several pieces were grozed, or cut for fitting into a window.

In form these early medieval windows varied considerably: the actual window openings were at first filled with thin sheets of marble, alabaster, gypsum, or even wooden boards, which were pierced with holes, coloured glass being inserted into these holes. In addition to glass, other materials were used for the same purpose, thin strips of alabaster set in bronze frame being not uncommon. This form, called a "mosaic" window, persisted even in western Europe into the Romanesque period, and 11th-century examples are found in Italy in the Cathedral of Torcello near Venice and in the Church of S. Miniato at Florence.

Early leaded windows

Leading was possibly used to hold together the pieces of glass in a window opening contemporaneously with the above early methods. Leading that may have been used in window glazing dating from the 4th century has been uncovered in excavations. The earliest example of a leaded window design was a small panel (destroyed in 1918) in the church at Séry-les-Mézières, northwest of Reims in France, probably of the 9th century. It appears certain that, as at Séry-les-Mézières, many of these early windows contained coloured glass arranged in comparatively simple decorative designs, with little use of the painted design.

There is no documentary evidence even to suggest the existence of pictorial windows until the 9th century, when several rather vague references testify to the appearance of figures in German and French glass. In the 10th-century history of the Church of Saint-Remi at Reims, it is stated that the windows contained various stories.

The fragments of what may be the earliest pictorial window extant were excavated at Lorsch in Germany. It was possible to reconstruct a head of Christ, which shows some stylistic affinity with Carolingian manuscript paintings and probably dates from the 9th, 10th, or 11th century. The earliest complete pictorial windows extant are those containing five figures of prophets in the Cathedral of Augsburg in Germany, belonging to the beginning of the 12th century (Figure 205, left).

In Carolingian and early Romanesque architecture the window openings, partly for structural reasons, were small and few in number. Polychrome decoration was naturally concentrated on the large mural areas and the vaults rather than on the windows. The development of late Romanesque and Gothic architecture brought a new emphasis on fenestration and openness. It was then that pictorial windows of stained glass became a major art form and in northern Europe the most important single element in church decoration.

Although the pictorial stained-glass window is normally regarded as the invention of and indigenous to western Europe, where its development can be followed with reasonable coherency from the beginning of the 12th century onward, there is still much that is obscure about its earlier evolution. The discovery in Istanbul of stained-glass windows, apparently deriving from a tradition independent of that in the West and datable to before 1136, adds to the complexity of the fragmentary evidence already cited.

12TH CENTURY

It is probable that many of the early stained-glass windows of the 12th century displayed a single monumental figure, like those depicted on the windows in the cathedrals of Augsburg and Canterbury or like the well-known Virgin and Child known as "La Belle Verrière" at Chartres. The most important feature of the 12th century, however, was the development of the narrative window, consisting of a series of medallions painted with pictorial subjects. This type of window was, so far as is known, first used extensively between 1140 and 1144 at the Abbey of Saint-Denis near Paris. A secondary but significant development of the second half of the century was the use of allover decorative patterns, or diapers, on the groundwork adjacent to the figures. This design device was probably more common at first in Germany than elsewhere, and an early example is in the Jesse tree window (*c.* 1170–80) at Frankfurt am Main, now in the city's Städelsches Kunstinstitut.

Regional schools of glass painting in France

France. By the 12th century the production of stained-glass windows in northern Europe was considerable, and regional schools begin to be discernible, especially in France, Germany, and England. In France a number of important regional schools of glass painting emerged, one of the earliest of which was in the west. The most important works of this group include the Ascension window (*c.* 1145) in Le Mans Cathedral and the Crucifixion window (*c.* 1165) in Poitiers Cathedral. In the northeastern region of Champagne appeared another quite distinct group, whose best work is found in the Redemption and St. Stephen windows (*c.* 1150–60) in the cathedral at Châlons-sur-Marne, together with the important later windows (*c.* 1190) at Saint-Remi in nearby Reims, whose stately figures indicate that Romanesque monumentality has already begun to be tempered by the less austere, less rigorously formal mode of the Gothic.

The most important workshop in the Île-de-France region around Paris was connected with the rebuilding of the choir of the Abbey of Saint-Denis. Only fragments of these windows are left, but the three windows (*c.* 1150–55) of the west facade at Chartres are later products of the Saint-Denis workshop and are a summation of all that is most uniquely Romanesque in stained glass.

The stylistic antecedents of these schools are difficult to pinpoint. The Saint-Denis-Chartres group has certain similarities to north French manuscript paintings that are not precisely dated, and the problem is further complicated by Abbot Suger's recording that the glaziers employed at Saint-Denis came from many differnt regions. The strongly Romanesque character of the Le Mans Ascension window, its general composition, and the particular stylization of drapery forms is similar to earlier manuscript paintings from western France. The Crucifixion window at Châlons-sur-Marne, on the other hand, has precedents in general arrangement in Ottonian manuscript painting and is also closely related in style and composition to the contemporary Mosan school of metalwork centred in the valley of the Meuse River. The similarities between the two are so marked that it is not impossible that the artist worked in both mediums.

12th century German glass

Germany. There is less 12th-century glass extant in Germany than in France. The outstanding example of German stained glass of the first half of the century is the series of five prophets (*c.* 1125) in the Cathedral of Augsburg (Figure 205, left). These hieratic figures have the monumentality of design, rigidly frontal and schematic, characteristic of Romanesque art. The bold use of ruby, green, yellow, and violet glass is completely alien to contemporary French developments. In the second half of the century, art in northern Europe generally, and perhaps more so in Germany, was influenced by Byzantine models. An example is the "Moses and the Burning Bush" window now in the Städelsches Kunstinstitut at Frankfurt am Main or the Magdalen (*c.* 1170) from the church at Weitensfeld, near Klagenfurt, in Austria.

England. England has only fragmentary remains of 12th-century glass. The nave clerestory windows in York Minster contain some reused panels from a series of narrative windows, one of which depicted the life of St. Benedict (*c.* 1140–60). Another panel, a single figure of a king from a Jesse tree, shows some affinity in style with the glass at Saint-Denis and Chartres but is probably later in date (*c.* 1190). The outstanding survival from the end of the century is the splendid series of figures representing the descent of Christ from Adam, made for the choir clerestory windows (*c.* 1178–1200) of Canterbury Cathedral, which resemble the "Prophet" windows in Saint-Remi at Reims. Their features show a new humanism, and there is a sense of movement, even tension, in their bodies and draperies, comparable to contemporary English manuscript painting.

13TH CENTURY

The grisaille window

A significant feature of the 13th century was the development of the grisaille window, composed largely of white glass, generally painted with foliage designs, and leaded into a more or less complicated geometric pattern (Figure 208). This type of design was employed partly as a means of introducing a larger amount of light and partly because it was considerably cheaper than coloured glass. The combination of grisaille glass and coloured subject medallions, or figures, however, disrupted the monumental overall unity, which is a feature of a window composed entirely of coloured glass, by allowing the penetration of pure light. This change had an important effect on style; the painted design became more linear and refined, the scale more

By courtesy of the Hessisches Landesmuseum, Darmstadt, Germany

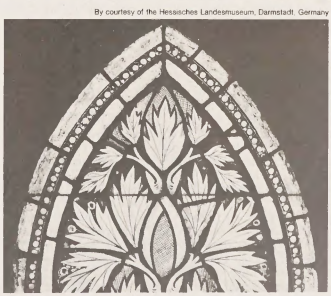


Figure 208: Grisaille panel from the Cistercian church at Altenberg, Germany, 13th century. In the Hessisches Landesmuseum, Darmstadt, Germany.

broken and delicate. Although the combination of grisaille and medallions, or figures, is not unknown in the early part of the century, it is more common in the second half, particularly in France and England.

The movement toward humanism, partly inspired by St. Francis of Assisi and his teaching, was accompanied by a related tendency toward naturalism discernible in the visual arts in the later 13th century. The conventional formalized foliage designs of the 12th and earlier 13th centuries gave place to more natural plant motifs of oak, vine, and maple, and these break out of the formal patterns and coil with a more organic natural movement.

Styles in 13th-century French glass

France. In France, where surviving material is most extensive, the various regional schools are mostly a natural development of their immediate predecessors. There is no radical change in style or technique during the first quarter of the century. In western France the severe Romanesque style was softened and refined, as seen in the Saint-Vital window (*c.* 1200) at Le Mans Cathedral, the Saint-Martin window (*c.* 1210) at the Cathedral of Angers, and the slightly later windows of Abraham, Lot, and Joseph at Poitiers Cathedral. Another distinct workshop, centred at Lyon and responsible (*c.* 1215–20) for the apse windows of Lyon Cathedral, is characterized by a strong Byzantine influence, particularly in iconography. An important workshop in Champagne had already produced in the late 12th century the clerestory windows of Saint-Remi at Reims that foreshadowed the mature Gothic style, while later works of this atelier can be found in the clerestory windows (*c.* 1235) of Reims Cathedral and the choir clerestory windows in the Cathedral of Troyes. In the Île-de-France and eastern France the situation is complicated by the almost complete loss of the later 12th-century works. The north rose window (*c.* 1200–05) of Laon Cathedral is stylistically related to the contemporary sculptures of the facade and to manuscript painting such as the Ingeborg Psalter (Musée Condé, Chantilly). The work of this atelier is extremely distinguished, with an elegance and purity of style and a knowledge of classical art that transcend most of its contemporaries. The rose window (1231–35) of Lausanne Cathedral in Switzerland was made by a wandering artist from Picardy, Peter of Arras, and is related in style and iconography to the Laon workshop.

The most extensive glazing program of the first half of the century was at Chartres Cathedral (Figure 207, right), a tremendous enterprise that brought together various workshops from different regions. The stylistic interactions between the different workshops resulted, particularly in the second quarter of the century, in a more general similarity of style between the various regional workshops. Contemporary with Chartres are the windows (*c.* 1214–35) of the east end of Bourges Cathedral, while four windows (before 1225) at Sens Cathedral are related in composition and ornament to Chartres and Bourges and also have certain stylistic affinities with the contemporary windows at Canterbury Cathedral.

Considerable activity was also centred in the Paris area during the second quarter of the century. The major monument of the period is the Sainte-Chapelle, which was built in Paris between 1243 and 1248. Forming what amounts to a continuous wall of 50-foot- (15-metre-) high stained glass around three sides of the chapel, it contains the most extensive narrative cycle ever produced in this medium, numbering 1,134 scenes in 15 windows. Stylistically these windows are closely related to Parisian court art of the same date, and their influence can be seen in the later windows at Le Mans Cathedral (about 1254–60) and in the Cathedral of Tours (1245/55–76).

England. The only extensive remains of 13th-century glass in England are found at Canterbury Cathedral, where the 12 Theological windows were produced in about 1200 and the windows relating to St. Thomas Becket in about 1200–30. Lincoln Cathedral retains impressive fragments of a series of windows made between 1200 and 1220, but it is impossible fully to appreciate either of these series without making comparison with French work. There are important similarities between Canterbury and Sens, on the one hand, and probably between Lincoln and Paris, on the other. The glaziers, however, were probably English with a close acquaintance with French models.

Features of 13th-century German glass

Germanic countries. Thirteenth-century stained glass in Germanic countries, however, was comparatively uninfluenced by French models. It is more turbulent in design, with agitated draperies, expressive faces, and a complicated ornamented character, particularly in the backgrounds. There were many distinct regional schools, among which Cologne was an important centre. The full-length figures of saints and the Legend of St. Kunibert window (*c.* 1220–30) in the Church of St. Kunibert at Cologne have elaborate geometrical frames around the figures and scenes that are without parallel in French art. These frames are a typical feature of German work; they occur again in later work of this school in the St. Nicholas window (*c.* 1240–50) at Büchen, west of Hamburg, and elsewhere. Another workshop, which produced the Jesse tree window (*c.* 1225) in the Cathedral of Freiburg im Breisgau, shows a marked affinity with the slightly earlier enamel and metalwork produced by Nicholas of Verdun and his circle (active 12th–13th century), while the remains of a group of windows (*c.* 1230) made for the Franciscan church at Erfurt constitute a particular group strongly indebted to Byzantine models. It appears that after the completion of the Erfurt windows, this workshop, or at least some of its members, went to Assisi in Italy and also to Gotland in Sweden. The most outstanding glass of the mid-century is a related series of windows in the cathedrals of Naumburg, Strasbourg, and Frankfurt. The Naumburg window of Holy Knights and Virgins can be contrasted with the delicate mannered style of Parisian court art.

Early examples of stained glass in Italy

The earliest Italian examples of stained glass were the three windows executed by German craftsmen of the Erfurt school in the apse of the upper church at Assisi between 1230 and 1240. At the end of the 13th century, native designers and craftsmen produced windows. The oculus window (*c.* 1288) in the apse of Siena Cathedral has been attributed to Duccio di Buoninsegna (died 1318 or 1319), and the style of the painter Guido da Siena (active *c.* 1250–75) seems evident in the Madonna and Child panel of the Sanctuary of the Madonna della Grotta near Siena.

EARLY 14TH CENTURY

Stained glass of the first half of the 14th century is everywhere distinguished by an insouciant fairy tale quality and a languorous charm sometimes tinged with pathos. Regional differences, however, persisted—the gentle reserve and earthy lyricism of the English; the virtuoso painting and exquisite drolleries of the Norman-French; and the full green-, gold-, and russet-dominated palettes of the German windows. The full flowering of the Gothic style side by side with the beginnings of stylistic developments that were to culminate in the Renaissance characterized the aesthetic nature of the early 14th century. The new movement toward the representation of volume and spatial depth, by means of modelling and perspective, had its origins in Flemish and Italian painting. That the glass painter was quickly influenced by this new style is seen, for example, in the St. Anthony window in the lower church of S. Francesco at Assisi, Italy. North of the Alps the earliest extant manifestation of this new interest in perspective and modelling, based on Italian models, occurs in the chancel windows (1325–30) of the Habsburg expiatory church at Königsfelden, near Brugg, Switzerland. The knowledge of Italian models spread quickly and extensively and can be seen in France in the windows (1325–39) at Saint-Ouen in Rouen and those made about 1330 at Évreux Cathedral. In the Germanic lands proto-Renaissance spatial illusionism influenced the transept windows at Augsburg Cathedral and the east window (*c.* 1340) of Vienna Cathedral, while the earliest remaining example in English glass painting is probably the nave windows (*c.* 1330–35) of Stanford-on-Avon in Northamptonshire, England.

In the early 14th century the third dimension in canopies was still highly untheoretical and largely governed by considerations of pure design. The practice of representing a figure beneath an architectural canopy was an established convention of the 13th century, particularly used

in clerestory windows. In earlier examples the canopy plays a comparatively unimportant part in the total design, but by the end of the 13th century, although still two-dimensional, it had become more elaborate and is an important ornamental feature of the windows of Merton College, Oxford. In German and Austrian windows the canopy work is often elaborate and complex in its spatial organization; examples are found at Vienna Cathedral (*c.* 1340) and Erfurt Cathedral (*c.* 1360–70).

Currents in France and England

The art of glass painting, however, did not respond equally to these new influences emanating from Italy. It subdivides itself into two groups, of which France and England together make up one, characterized by its resistance to Italian influences. The use of perspective was purposefully restrained, so that the essential overall surface unity of the design was not violently upset, for the use of flat, patterned diaper grounds effectively counterbalanced the suggestion of spatial effect. In the first half of the century the most important work in France is found in the region of Normandy, especially in the choir windows (*c.* 1330) of Évreux Cathedral and those dating around 1325 to 1339 at Saint-Ouen in Rouen. The English glaziers of this period were extremely prolific, with Oxford, Coventry, and York as important regional centres. The nave windows of York Minster were made between 1300 and 1338 and are the largest single enterprise of this period in England. It appears probable that some of the later glass at York (*c.* 1350–70), now distributed over the windows of the choir clerestory, was the work of an imported French glazier, probably from Rouen. A flourishing school in western England, whose best work is found at Wells Cathedral and Eaton Bishop in Herefordshire, shows some affinity with German work. The best French and English work, however, has a lightness of colour and graphic refinement that is enhanced by an extensive use of yellow stain. After about 1300 the geometric grisaille glass gave way to simpler diamond-shaped pieces, painted with delicate trails of foliage and leaded together to give the effect of trellis work.

Three-dimensional style of the Germanic countries

The second group, which might be termed "Germanic"—as it embraces Germany, Bohemia, and Austria—displays a much more three-dimensional style; the colours are deeper and more saturated, the compositions are more complex, both on the surface and in depth, and the canopy designs particularly are often complicated essays in perspective; the figures are shorter in proportion and their volume more accentuated. It is brusque, almost harsh, contrasting strongly with the elegance of French and English work. All of these traits can be seen, for example, in the panels (*c.* 1350) from Strassengel, near Graz, Austria, now in the Victoria and Albert Museum, London; the earlier windows (*c.* 1350–60) at St. Maria-am-Gestade in Vienna; and at Erfurt Cathedral (*c.* 1360–70). A particular trait of this Germanic group, of which Erfurt is a good example, is a tendency to extend a single composition across the main lights of a window, ignoring the natural divisions of the stonework. The reason for this is partly architectural: the window lights are comparatively much narrower and taller than those in French or English windows.

LATE 14TH, 15TH, AND 16TH CENTURIES

The International Style

The arts from about 1380 to about 1430, which are frequently grouped together under the title of the International Gothic Style, belong essentially to an era of court art inspired by the patronage of kings, the nobility, and the higher orders of ecclesiastics. Surviving windows are extensive, and interactions between the various centres of patronage, complicated by family alliances and the exchange of works of art and artists, are particularly complex. The various national and regional styles can still be distinguished in glass painting, but there is a general tendency toward a mannered, extremely sophisticated elegance of style, sometimes verging on the precious, combined with an interest in portrait realism.

If the genius of the 13th-century stained glass lay in its epic sense of monumentality and that of the early 14th century in its warmth of human feeling, that of the late 14th and early 15th centuries is far more difficult to characterize in a phrase. The style of glass painting became at once more corporeal and more introspective. Figures are rendered with far more attention to individual human traits yet are akin in their majesty to the great prophets in the 13th-century windows. The figures painted between 1384 and 1392 by Hermann von Münster, for the west window of Metz Cathedral in France, and those painted in about 1400 by Thomas of Oxford for the chapel of William of Wykeham's College at Winchester, England, seem both newer and older than their immediate predecessors—newer in that they are beginning to bear the signature of a personal style, older in that they recall the grandeur of an earlier, essentially collective expression that now seemed everywhere to be giving way to arid cameo-like refinements on the one side and an earnest rusticity on the other.

The glass painting of this period is of high quality. The most significant examples in France are the "Royal" windows (*c.* 1395) at Évreux Cathedral. One of the most ambitious works of the period is the great east window of York Minster, made between 1405 and 1408 by John Thornton of Coventry.

In Germany and Austria there was a softening and refining of the "plastic style" of the earlier part of the century. In Austria an important atelier associated with the court produced the window (*c.* 1386–95) in the church of St. Erhard in der Breitenau and later the series of windows (*c.* 1400) at Viktring, and for the Freisingerkapella at Klosterneuburg (*c.* 1410). Germany contains a large amount of work of this period. At Erfurt Cathedral the window given by Johann von Tiefengruben (*c.* 1400) is lighter in tone and has a more refined pictorial style than the earlier windows there. The series of windows at Rothenburg, which were probably made in about 1400, should be noted, together with the cycle of windows (*c.* 1420–30) for the Bessererkapelle at Ulm Cathedral, which show an admirably refined technique.

The period 1430–1550 saw not only the decline of the Gothic style and the establishment of the new Renaissance style but also the beginning of the transformation of the art of glass painting from a significant means of artistic expression into a hybrid art form: the translucent emulation of fresco and easel painting.

The International Gothic Style continued to influence glass painting during the first half of the 15th century but to a lesser degree. Its mannered elegance and extravagant costumes can still be seen in France in the two rose windows (*c.* 1440) at Le Mans Cathedral and also the rose windows (1441–42) at Angers Cathedral. In England this aesthetic is continued in the east window (*c.* 1423–39) of the priory in Great Malvern (Worcestershire), the chapel windows (1441–47) of All Souls College at Oxford, and in the windows made by John Prudde, the king's glazier, for the Beauchamp Chapel of St. Mary's in Warwick, which were commissioned in 1447. The work of Germanic glass painters provides outstanding examples, particularly the window (after 1424) from St. Lambrecht, now in the museum at Graz, Austria, and the charming Alsatian window of St. Katherine (*c.* 1425–50) at Sélestat, France.

The Flemish style

There were, however, important new influences affecting the styles of glass painting. In northern Europe during the first half of the 15th century a flourishing school of painting emerged in Flanders. Although the origin of the Flemish style is partly to be found in the International Gothic, a more realistic manner of representation and a detailed awareness of actuality developed that is almost the antithesis of the mannered sophistication and the essentially unrealistic world of that style.

It was impossible for the glass painter to participate fully in the new realism. Glass is a translucent material; the passage of light through it is alone sufficient to create a feeling of unreality. Furthermore, although the panel or mural painter could use the new discoveries of linear perspective to heighten the sense of reality, the glass painter had to contend with the presence of the leading line, which emphasizes the surface plane. This creates a tension and a sense of ambiguity between the actual surface and the illusion of depth. Attempts were made to resolve the problems, but with little success; one result was that the lead line became increasingly divorced from the design

instead of being an integral part of it. The increasing use, from the mid-16th century onward, of vitreous enamel pigments had the effect of accelerating this process.

The full flowering of the Renaissance style in Italy and the intense interest in classical art that it stimulated resulted, in about 1500, in a profusion of ornamental details, borrowed from the formal language of classical art, in contemporary window designs. In addition the glaziers at this time often drew inspiration from contemporary engravings, particularly those of Albrecht Dürer.

The period 1450–1550 saw no decline in demand for stained glass. One of the most productive and influential areas at this time was Flanders. The six windows (*c.* 1475–1500) of the church of St. Gommaire at Lier and the Virgin window (*c.* 1482) at Anderlecht, although restored, have original portions of excellent quality and close affinity with contemporary Flemish painting.

English stained glass in the 15th century becomes more intimate, more anecdotal. It is less a cathedral art than an art of parish churches and is addressed less to an assembled ecclesia than to the individual believer. At its best it achieves a quiet intensity in the woodcut-like figures in East Harling church (Norfolk), a pathos in the Long Melford church (Suffolk) "Pieta," and a sheer power of expression in the Clavering church (Essex) "Martyrdom of St. Catherine."

The trading and political links between Flanders and England in the second half of the 15th century encouraged the influx of Flemish works of art and artists into England. A number of Flemish glaziers established themselves in Southwark, London, and some even assumed English names. The early 16th century was marked by a series of disputes between the London Guild of Glaziers and the foreigners. The latter were particularly patronized by the court and the more wealthy merchants. The two outstanding monuments of this imported style are the windows (*c.* 1480) of Fairford church (Gloucestershire) and the windows (1515–31) at King's College Chapel, Cambridge, many of which were probably designed by the Antwerp painter and engraver Dirk Vellert.

New realism in French glass painting

The spread of the new realism can be traced in French glass painting. The interactions with Flemish glaziers can be seen, for example, in the realism of the windows (*c.* 1430–40) from the Duke of Burgundy's chapel at Dijon, now in the Victoria and Albert Museum, London, which are the work of Flemish glaziers, and also the Jacques Coeur window (Figure 209) at Bourges Cathedral, which is markedly Flemish in style. Native French glaziers were extremely prolific at this time, and Normandy and particularly the city of Rouen contain an incomparable display of windows produced by a large number of distinctive workshops. The leading figures of the first part of the 16th century were Arnoult of Nijmegen (*c.* 1470–1540) and the Le Prince family at Beauvais. Arnoult of Nijmegen worked in both Flanders and France. His most important works are the windows he executed between 1490 and 1500 in Flanders for Tournai Cathedral and the Jesse window (1506) in Saint-Godard at Rouen, which is one of the most impressive examples of glass painting of the period. The works by the Le Prince family are equally impressive, particularly the Jesse tree window (Figure 207, left) in Saint-Étienne at Beauvais.


Photophot—Ziolo

Figure 209: *Continuous narrative in four windows.* "The Annunciation," the Jacques Coeur window, *c.* 1450. In Bourges Cathedral, France.

In Germany the later 15th century is dominated by the prolific activity of Peter Hemmel von Andlau (active 1430–1500) and his workshop. Examples of his work were done for St. Wilhelm's Church in Strasbourg.

In Italy a remarkable ensemble of stained-glass windows was created in the first half of the 15th century for the cupola of Sta. Maria del Fiore in Florence by some of the greatest masters of early Renaissance art: Ghiberti, Donatello, Castagno, and Uccello. Domenico Ghirlandajo created in the second half of the century windows for the Florentine church of Sta. Maria Novella. The late 15th and early 16th centuries are mainly associated in Italy with the name of Guglielmo de Marcillat (1467–1529), a Frenchman whose works display a thorough mastery of technique. His finest windows are at Arezzo Cathedral. The building of Milan Cathedral caused an important school of glass painting to develop there, and the work of Conrad Munch, a German from Cologne, and Nicolo da Varallo is noteworthy. The Milan school continued in full activity during the 16th and 17th centuries.

In Spain—especially at Seville, Leon, and Avila—there are some good 16th-century windows. They are in all cases the work of imported Flemish glass painters.

17TH AND 18TH CENTURIES

The spatial illusions of Baroque paintings were beyond the limitations imposed by the stained-glass medium. The glass painter of the 17th and 18th centuries found himself reduced to completing the cycles of stained-glass windows in medieval churches or to creating contemporary art for an architecture with no artistic affinity with traditional stained glass. The most interesting development in the late 16th and early 17th centuries was the intimate and portable heraldic panel, which became fashionable to hang in domestic windows, particularly in Switzerland, the Low Countries, and Germany. These panels, seldom more than two feet high, are the glass painter's showpieces; they complete the divorce between stained glass and architecture.

The fashionable heraldic panel

Painting glass with vitreous enamels in the 17th and 18th centuries led to the final decline of the art of stained glass. In the St. Janskerk windows at Gouda, Holland, painted by the brothers Wouter and Dirk Crabeth at the end of the 16th century, and in the works (1620–40) of Abraham and Bernard van Linge, the realization of the window as a translucent canvas painting is complete. Abraham van Linge's windows painted in 1630 to 1640 for Christ Church Cathedral at Oxford are an excellent example of the destruction of the lead line as an integral part of the design. The leading simply holds together the square sheets of glass: the effect is the same as looking at a picture set behind a rectangular grid. This type of design was continued by English glass painters such as Henry Gyles and the Price and Peckitt families, all of York, Francis Eginton, and Sir Joshua Reynolds, who designed in 1778 the west window for New College Chapel, Oxford.

19TH CENTURY

The Gothic revival that came as an offspring of the Romantic movement of the late 18th and early 19th centuries represents the beginning of a revitalization of the art of stained glass. The revival of interest in Gothic art stimulated an interest in both the technique and history of medieval glass painting. The pioneer figures in this field were E. Viollet-Le-Duc in France and Charles Winston in England. Winston was a lawyer and antiquarian who associated with various London glaziers and, with the technical help of James Powell and Sons, brought about a considerable improvement in the technical quality of coloured glass. In 1847 he wrote the first comprehensive study of the medium. The experiments were continued by W.E. Chance, who first successfully produced "antique" glass in 1863.

Revival of medieval glass painting

In the first half of the 19th century the styles and methods of the early Gothic period were reconstructed, but without much aesthetic appreciation of medieval art. Much of the work was stereotyped and mass-produced, particularly in Germany, and varied considerably in technical quality. The latter part of the century is dominated in England by Edward Burne-Jones and William Morris. Burne-Jones provided the designs and Morris adapted them to the medium of stained glass. In windows by them the lead line is once again treated as an integral part of the design, as seen, for example, in the windows for Christ Church at Oxford (1874–75 and 1878), Salisbury Cathedral (1879), and Birmingham Cathedral (1897). In the U.S. the works of John La Farge and Louis Comfort Tiffany were influential in creating an American interest in stained glass. Although the style and sentiment of 19th-century work has not been much in favour in the 20th century, the period had great historical significance in the revival of the basic technique of making stained glass.

Art Nouveau designers used stained glass decoratively for making such objects as lampshades and light fixtures, and turn-of-the-century architects increasingly employed stained glass as an integral element in wholly modern architectural settings: Victor Horta in his Hotel Solvay (1895–1910), Brussels; Antonio Gaudí in his Chapel of Santa Coloma de Cervelló (1898–1914) in the Güell Colony near Barcelona; Charles Rennie Mackintosh in the Willow Tea Rooms (1904), Glasgow; and Frank Lloyd Wright in the Coonley House (1908), Riverside, Illinois, and the Unity Church (1906), Oak Park, Illinois. These windows and panels clearly mark the beginnings of an authentically modern stained glass, despite their strictly ornamental intent.

20TH CENTURY

Three interrelated creative currents can be discerned in the development of 20th-century stained glass. First, a significant number of architects, following the lead of their turn-of-the-century predecessors and taking advantage of the new systems of fenestration made possible by modern structural engineering, have continued to discover many new ways of using stained glass. Second, especially in post-World War II France, several major easel painters turned their attention to stained glass, infusing it with many new and powerful images. Third, during the 1950s and 1960s Germany produced the first authentic school of stained glass since the Middle Ages, dedicated to exploiting the unique technical and expressive resources of the medium.

Although the bulk of significant 20th-century stained glass belongs to the period after World War II, earlier experiments, especially in France and Germany, suggested the possibilities that could be creatively explored. In Auguste Perret's church of Notre-Dame (1922–23) in Le Raincy, near Paris, the entire wall surface becomes a geometric grillwork of coloured glass by the Symbolist painter Maurice Denis. In 1930 the Dutch-born artist Johan Thorn Prikker completed a cycle of windows for the Romanesque Church of St. George in Cologne in which lead lines are used with a graphic eloquence and deep smoldering colours with a monumental gravity that have no parallel even in the greatest medieval windows.

After 1946 there was an unusual burst of activity. In Le Corbusier's Notre-Dame-du-Haut (1952–55) at Ronchamp, northwest of Dijon in France, the massive south wall, 12 feet (2.72 metres) thick at the base and five feet at the top, is dramatically punctuated with a series of crude, yet remarkably effective, stained-glass windows through which shafts of light fairly explode into the church. Simultaneously, in Dominikus Böhm's and Heinz Bienefeld's Church of Maria Königin (1953–54) in Cologne-Marienburg an entire sidewall of the church is conceived as a diaphanous veil of silvery gray stained glass that half reveals and half conceals the parklike grounds outside with equally dazzling effect. In Wallace K. Harrison's First Presbyterian Church (1958) in Stamford, Connecticut, the whole central section of the church is nearly engirdled with slab glass and concrete. In all of these structures, different as they are in nearly every other respect, stained glass is seen boldly exploited once again not merely for its local ornamental quality but as a major, integral atmosphere-creating element.

Work after World War II

The most seminal contributions of the School of Paris painters to the art of stained glass were Henri Matisse's Chapel of the Rosary (1948–52) in Vence and Fernand Léger's windows for the Sacré-Coeur (1950–52) in Audincourt. Both are by artists whose manner was rather directly translatable into stained glass. It was but a comparatively short step from Matisse's large coloured-paper collages to the disarmingly simple decorative windows in Vence, but the way Matisse used them to create an enchanting play of colour in the chaste white space of the chapel is masterful. And it took the boldly emblematic style of Léger to reveal the true expressive potentialities of slab glass and concrete. A third important work of this period is the long friezelike window created by the sculptor Léon Zack for the Church of Notre-Dame-des-Pauvres (1955) in Issy-les-Moulineaux, remarkable for its daring sequence of colour harmonies and delicate lead line motifs reminiscent of the art of Paul Klee. The stained-glass windows of Georges Braque, Jacques Villon, Georges Rouault, Marc Chagall, and Alfred Manessier are also noteworthy if less authoritative in their handling of the medium.

In Germany such distinguished prewar church architects as Dominikus Böhm and Rudolf Schwarz and the stained-glass artist Anton Wendling were able to resume careers interrupted by the Nazi era and to set the course for a whole new generation of stained-glass artists, especially in the Rhineland. Inspired by the example of Thorn Prikker, these artists have continued to explore the unique qualities of stained glass—the special refractory properties of opal-flashed antique glass, the graphic potentialities of the lead line, the bold effects of texture and relief that had become possible with slab glass and concrete—and to create a whole gamut of strange brooding colour harmonies the like of which had not been seen in stained glass since the Augsburg prophets. Among the more important works of this Rhenish school are Georg Meistermann's windows for the Dom Sepulchur (1957) in Würzburg and his complete ensemble of windows for the 15th-century church of St. Matthew (1964) in Sobernheim; Ludwig Schaffrath's cycle of modern grisaille windows for the cloister (1962–65) in Aachen, his high triple-gabled window walls for the transepts of St. Peter's Church (1964) in Birkesdorf, near Düren, and his powerfully iconic and technically innovative slab- and rod-glass sanctuary window in St. Matthew's Church (1966–67) in Leverkusen; Wilhelm Buschulte's unusually rich colour harmonies in his cycle of nave windows for the Cathedral of Essen (1964) and the choir of the Church of SS. Peter and Paul (1967) in Wegsburg, near Mönchengladbach; and Johannes Schreiter's almost monochromatic Abstract Expressionist windows for the Church of St. Margaret (1961) in Bürgstadt. Trained once again to work of the scale of the cathedral windows and to develop their art in accordance with its own intrinsic potentialities, such artists have been collaborating with some of the best architects in Germany to create the most impressive body of stained-glass windows since the Middle Ages.

(R.So./Ed.)

THE HISTORY OF GLASS DESIGN

From very early times glass has been used for various kinds of vessels, and in all countries where the industry has been developed glass has been produced in a great variety of forms and kinds of decoration, much of it of great beauty. For the composition and properties of glass and the manufacture of various glass products such as glass containers, window glass, plate glass, optical glass, and glass fibres, see INDUSTRIAL GLASS AND CERAMICS.

This section is concerned with the aesthetic or artistic aspect of glass.

Antiquity and the Middle Ages

EARLY GLASS

It is not certain in which of the civilizations of the ancient Near East glass was first made. The earliest wholly glass objects from Egypt are beads dating from some time after *c.* 2500 BC. A green glass rod found at Eshnunna in Babylonia may go back earlier, possibly to 2600 BC. A small piece of blue glass found at Eridu dates from before 2200 BC. The manufacture of glass vessels, which may have begun slightly earlier in Mesopotamia, was carried to a high point of excellence in Egypt during the 18th dynasty (*c.* 1490 onwards). These vessels are distinguished by a peculiar technique: the shape required was first formed of clay (probably mixed with dung) fixed to a metal rod. On this core the body of the vessel was built up, usually of opaque blue glass, on which, in turn, were coiled threads of glass of contrasting colour. The threads were pulled alternately up and down by a comb-like instrument to form feather, zigzag, or arcade patterns (Figure 210). The threads—

Egyptian technique

By courtesy of the trustees of the British Museum

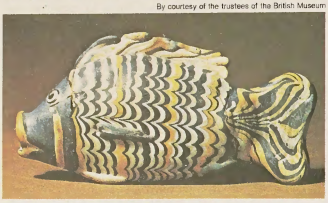


Figure 210: Fish of core-made glass with "combed" decoration, Egyptian, New Kingdom, 18th dynasty (*c.* 1363–46 BC). In the British Museum. 0.141 m × .069 m.

usually yellow, white, or green in colour, and sometimes sealing-wax red—were rolled in (marvered) flush with the surface of the vessel. Finally, if desired, handles—often of translucent glass and sometimes of patterned "canes"—were added. The vessels were nearly always small, mainly for unguents and the like. Occasionally glass was decorated on the lapidary's wheel. Glass is known to have been made on the palace site of Tell el-Amarna, the residence of Akhenaton (reigned *c.* 1379–62 BC), and the number of fragments found in and near the palace of Amenhotep III (reigned *c.* 1417–1379 BC) at Thebes suggests that it was made there also. This palace activity seems then to have died down and after the 21st dynasty (1085–945 BC) to have ceased altogether.

In Mesopotamia the Nineveh tablets of the reign of Ashurbanipal (668–*c.* 626 BC) and the remains of glass in various forms excavated at Nimrūd (ancient Calah, Assyria) indicate that glassmaking was carried on there during the 8th to the 6th centuries BC. It is probable that certain vessels of palish-green or deep blue glass, cut from a solid mass as if from stone, are Mesopotamian and date from as early as the 8th century BC, as a dish from controlled excavations in Phrygia proves. A vase of this type, contrasting completely with the core-wound glass of Egypt, bears the cartouche (panel enclosing the name) of the Assyrian king Sargon II (reigned 721–705 BC), and it is probable that glass treated in this way was manufactured over a long period in Mesopotamia.

Glass was made in Greece in Mycenaean times (*c.* 1400–1200 BC) usually in the form of small molded architectural details. A few pieces suggest, however, that perhaps some vessel glass also was made by the Egyptian technique, though not in Egyptian forms. Other Aegean-area glass of this period may have been imported from Egypt.

In general, glass of the earlier half of the 1st millennium BC is scarce and displays little homogeneity. From the 6th century BC, however, glass begins to appear in great quantities once again, particularly on the Greek-inhabited islands of the Aegean, in Greece itself, in Italy and Sicily, and even farther west. This contrasts with the meagre contemporary finds on Egyptian soil. The later glasses in the old Egyptian core-wound technique were probably made in Syria or some part of the Greek world. Such vessels were still small but differ in shape from the earlier Egyptian dynastic work. They were usually decorated with light-coloured threads on a dark, usually blue, ground (familiar from the Egyptian 18th dynasty), but a notable variation was displayed in pieces decorated with dark purple threads on a white ground. In the Hellenistic period (roughly from the 4th century BC) the shapes of glass degenerated. The technique of decoration, however, remained the same; new colour combinations were used, and indeed these combinations continued into the era of blown glass.

THE ROMAN EMPIRE

In Egypt during the Ptolemaic period (330–305 BC) Alexandria came to the fore in glassmaking. By about the 1st century BC, which saw the beginnings of glass as known today, it had become preeminent in certain glass techniques. Alexandria inherited and perfected the manipulation of coloured glass rods to make composite canes, which, when cut across, revealed a design (mosaic glass). Slices from such canes could be arranged side by side to produce repetitive patterns. When, as often happens, the cane slices show starry or flowerlike designs, the resultant glass is called millefiori ("thousand flowers"). An Alexandrian technical speciality more important for the future, however, was molding, glass being pressed into, or powdered glass melted in, a mold. A combination of this process with the millefiori technique produced bowls with variegated designs in infinite variety (Figure 211). Sometimes glass of various colours was irregularly compounded to give the effect of a natural veined stone; occasionally enclosures of gold leaf in the glass simulated the glitter of natural pyrites (aventurine glass). Bowls were often finished around the rim with a cordon made of a clear glass thread twisted with one of opaque white. Sometimes such cable threads were themselves coiled round and round from a centre to make a bowl of lacy appearance, with the opaque white glass threads apparently set in a clear colourless matrix.

Dominance of Alexandria in glassmaking

By courtesy of Victoria and Albert Museum, London



Figure 211: Footed bowl of pressed mosaic glass, probably Alexandrian, 1st century AD. In the Victoria and Albert Museum, London. Diameter 15.5 cm.

All these pieces might be finished with a fire polish by returning them to the furnace, but many mold-pressed glasses were, in fact, given a rotary polish, either by means of a spinning wheel fed with abrasives or by a process similar to lathe turning, in which the object spins and the tool is stationary. Similar equipment probably produced the numerous pieces that give every appearance of having been cut from a solid block of glass or at least from a thick, mold-pressed blank. Such pieces (usually flat dishes or two-handled cups) follow the contemporary forms of pottery and metalwork. Wheel engraving appears to have become an Alexandrian specialty around the 1st century

BC and probably continued so throughout the two succeeding centuries. Alexandrian wheel engravers produced not only massive cut shapes, but also intaglio (incised) and relief surface decoration, the latter by laboriously grinding back the surface of the glass to form a background for the design. Simple motifs such as lotus buds or lotus flowers were produced in this way and occasionally more elaborate figural compositions were also done. Other specialties attributed to Alexandria were enamel painting (pigments mixed with a glassy flux were fused to the surface of the glass vessel by a separate firing) and an extraordinary technique of sandwiching a gold leaf etched with a design between two layers of clear glass.

Introduction of blown glass

The most important innovation in the whole history of glass manufacture was blowing. Perhaps by a stroke of pure inventive genius it was perceived that glass on the end of a hollow metal tube could be blown into a mold as easily as it had theretofore been pressed in. The next stage was to use molds for forms, such as flasks, that could not be made by pressing. Finally, it was realized that the glass bulb on the end of the blowpipe could be shaped freehand to any form desired, and handles, feet, and decorative elements could be added at will. This liberating discovery, probably made during the 1st century BC, gave rise to the astonishing growth of the glass industry in Roman imperial times. In addition to the luxury vessels of types already described, which were produced with an elaboration of skill that astonishes and often baffles the modern technician, commercial containers in great variety were mass-produced in common greenish glass on a scale that was not matched until the 19th century.

The discovery of glass blowing may well be credited to the Syrian glassworkers, since the first mold-blown glasses bear the signatures of Syrian masters and since the readily ductile Syrian soda glass was especially apt for this purpose. Syrian glassworkers, however, seem to have migrated wherever demand promised a ready market, and some masters of mold blowing appear to have moved to Italy early in the 1st century AD; in the course of that century Italy became an important glass-producing area. Glass engraving especially seems to have flourished there and particularly one form of the art—grinding through an opaque white layer to a darker ground (cameo glass). The most famous example of this exacting technique is the Portland vase, in the British Museum, London. The capacity of the Italian glass craftsman to surpass all earlier masters in work of the most complex character is seen in the so-called cage cups (*diatreta*), on which the design—usually a mesh of circles that touch one another, with or without a convivial inscription—is so undercut that it stands completely free of the body of the vessel, except for an occasional supporting strut (Figure 212). These cups were made perhaps at Aquileia and date from the 3rd and 4th centuries.

Cameo glass engraving technique

Parallel to the pottery industry, glassmaking spread from Italy to northern Gaul, in particular to the valleys of the Rhône and the Rhine. In Britain the industry was probably not of great importance. The Rhineland, however, became one of the great glassmaking areas of the Roman world (partly, it is thought, because of successive migrations of Near Eastern workers) and, although Rhenish glass is always recognizably Roman, several types of decorated glass were specialities of the district. Glasses decorated in serpentine patterns by threads trailed on and then pressed flat and notched are perhaps the most important and typical (*Schlangenfadengläser*). A considerable school of glass engraving also seems to have flourished, probably around Cologne. Although some engraving shows an impoverished linear style eked out by lines scratched with a hard stone point, some is executed by means of wheels sufficiently thick to permit rounded cuts corresponding to the modelling of the human figure, and simulating it when the piece is seen against the light. Both types of decoration flourished in the 3rd and 4th centuries.

In Egypt in the later centuries of the Roman epoch glass was in frequent use for tableware, but artistic standards were not high. Plain dishes, cups, bowls, and lamps are characteristic; the glass of such tablewares ranges from an almost colourless "metal" (basic glass) of good quality to a greenish brownish substance full of bubbles and impurities. Decoration in this late period is mainly restricted to a few rough-cut lines, an occasional group of coloured glass blobs on the lamps, or a zigzag trail of glass thread running between the lip and the shoulder of a vase. In Syria during the same period, however, this trailing technique, which was particularly suitable to the ductile Syrian material, was carried to extreme lengths—threads circling the body or neck of a vessel, a profusion of zigzags, and fantastically worked handles.



Figure 212: An example of *diatreta* ("cage cups"), perhaps made for a Greek living in the Rhineland, 3rd to 4th century AD. In the Römisch-Germanisches Museum, Cologne. Height 12.1 cm.

By courtesy of Romisch-Germanisches Museum, Cologne

Syrian use of trailing technique

With the breakdown of the Roman Empire, glassmaking fared differently in different parts of the world. In the East, urban life continued relatively undisturbed, and glassmaking evolved in an unbroken progress into Islāmic times. In the northern provinces, however, glassmaking became an affair of small, often isolated, glasshouses working in the forests that supplied them with fuel. Relatively simple shapes were made of an impure greenish or yellowish material, and decoration was restricted to simple trails of thread. Considerable virtuosity, however, was displayed from *c.* 500 onward in the manufacture of the elaborate

By courtesy of the trustees of the British Museum



Figure 213: *Rüsselbecher* Frankish glass, probably 7th century AD. In the British Museum. Height 19.0 cm.

and fantastic *Rüsselbecher* ("elephant's trunk, or claw beaker") on which two superimposed rows of hollow, trunklike protrusions curve down to rejoin the wall of the vessel above a small button foot (Figure 213).

In the East, Syria appears to have continued its predilection for trailed and applied ornamentation. In Egypt the art of glass suffered a catastrophic decline; only small rough vessels of impure green or blue material were manufactured.

BYZANTIUM

In Byzantium itself the position of glassmaking is obscure. A distinction was made between *vitrarii* ("glassmakers") and *diatretarii* ("glass cutters") in edicts of Constantine the Great, Theodosius, and Justinian—suggesting that cutting played an important part in Byzantine glass decoration. This is borne out by the fact that cut glass made up the greater part of the glass that was brought back from the sack of Constantinople by the crusaders and placed in the treasury of St. Mark's in Venice. Apart from a few pieces of obviously Roman glass, presumably kept as heirlooms in Byzantium, these glasses are decorated either with tessellated (mosaic) patterns of overlapping round or oval facets or with round bosses in relief. These same two forms of cutting are observable in glass of the 5th century excavated at Kish in Mesopotamia; it is a fair assumption that Byzantine taste in glass, as in some of the other arts, was strongly influenced by the East. It is probable, however, that some enamelled and gilt glass also was made in the Byzantine provinces (*e.g.,* in Corinth), if not in Byzantium itself.

Glass from the sack of Constantinople

ISLĀM

In the 7th century the whole Near East was overrun by the Arabs, and a number of rival dynasties were established in different parts of the conquered territory. An Islāmic civilization developed comparable to the preceding area of Greco-Roman culture, and a distinctively Islāmic glass style evolved. Although often it is not possible to say where a particular glass was made, different parts of the Islāmic world seem to have shown predilections for one or another type of glassmaking. In Syria, pieces more or less heavily decorated with trailed threads or applied blobs and pieces blown in molds, patterned with ribs or other allover designs, were still made. In Mesopotamia, glassmaking and, in particular, engraving flourished, especially during the ʿAbbāsid dynasty (750–1258), and attracted many of the best artists in the Islāmic world. Not only were the earlier modes of facet and boss cutting continued, but (perhaps deriving from them) two splendid new styles were created, one of linear intaglio, the other of relief cutting (outlines were left in relief by cutting back the ground and were then enlivened by crosshatching). Bowls, bottles, and ewers of remarkable sumptuousness were decorated with forms of running animals and plant scrolls. The quantity of engraved glass of these types found in Persia suggests that such work was done there also.

In Egypt there was both innovation and, after the post-Roman period, a notable revival of earlier techniques. Among the innovations was the stamping of glass by means of tongs, one jaw of which was patterned. The technique also is found in other lands. One extension of it, by which a bottle's upper and lower halves, made separately in contrasting colours, were decorated by the tongs and then joined together, was probably a Syrian innovation. More important was the Egyptian invention of lustre painting. In its simplest form it consisted of painting with a pigment containing silver that when fired in a smoky atmosphere (*i.e.,* without oxygen) produced on the glass a thin, metallic film that varied in colour from pale yellow to brown. Intact bowls and a bottle decorated by this technique exist, but whole classes of much more elaborate lustre-painted glass are represented only by fragments. A very wide variety of sumptuous polychrome effects are represented, although many were probably not produced by lustre properly so-called. The technical processes by which these effects were achieved are not yet understood.

Egyptian tong stamping and lustre painting

Egyptian Islāmic revivals in glass included millefiori effects, found mainly in plaques for wall decoration, and white fern and feather patterns that were produced on dark glass vessels by combed and imbedded glass threads. Glass cutting was also practiced in Egypt, primarily for the production of deeply incised small perfume bottles of square sections, the bases of which were often cut into four tapering feet ("molar tooth" bottles). It seems probable that in Egypt was also perfected the techniques of gilding, decisive for the next phase in Islāmic glassmaking. In gilding, gold leaf is applied to an object that is then fired to fix the glass.

Glassworkers migrating from Egypt to Syria after the fall of the Egyptian Fāṭmid dynasty in 1171 may have laid the foundation of the Syrian art of enamelled and gilt glass. Although earlier phases of this art are incompletely understood, the first group of enamelled and gilt glasses seems to be one in which thick enamels are used (particularly white and turquoise blue), often in series of beadlike drops; this group is tentatively associated with the town of Raqqah in Syria. A similar doubt surrounds the origins of two broad families into which Syrian glass of the 13th century is divided. One, characterized by the use of thick, jewellike enamels, is connected with the town of Aleppo; the other, notable for its exquisitely painted small-scale figural decoration, is attributed to Damascus.

Syrian enamelled and gilt glass

Both cities were famous for their glass at this time, but it is uncertain what each produced. Wherever made, these two types of glass represent one of the highlights in the history of the art, whether one considers the rich green, red, yellow, white, and turquoise-blue enamels of the "Aleppo" group or the masterly red outline drawing of the "Damascus" group.

Toward 1300, Chinese influence, infiltrating by way of the Mongols and Tatars, makes itself felt in the decoration of these glasses (Figure 214), as is apparent in the series of great mosque lamps that then began to be inscribed with the names of rulers and great officers of state in Egypt. From a peak of excellence at the beginning of the 14th century a decline set in, greatly precipitated by the Mongol conqueror Timur's sacking of the chief Syrian cities at the end of the century.

Damascus fell finally in 1400, and it is recorded that the glassworkers of that city were carried into captivity in Samarkand. Nevertheless, some enamelled glass of in-



Figure 214: Bottle of enamelled and gilt glass decorated with Chinese motifs and an inscription in Kufic lettering praising an unknown sultan, Syrian, Mamlūk period, *c.* 1300. In the Victoria and Albert Museum, London. Height 43.5 cm.

By courtesy of Victoria and Albert Museum, London

ferior quality continued to be made in the 15th century, perhaps in Egypt. By the end of that century, however, there is evidence that mosque lamps were being made in Venice for the oriental market and the great Near Eastern tradition of enamelled and gilt glass was clearly moribund.

Mid-15th to mid-19th century

VENICE AND THE FAÇON DE VENISE

A glass industry was already established near Venice in the 7th century, and vessel glass was made there by the last quarter of the 10th century. In 1291 the glass furnaces were removed to the neighbouring island of Murano to remove the risk of fire from the city. Although Venice had constant contact with the East, there is no evidence that it was indebted to that source for its skill in glassmaking. Venetian enamelled glasses (Figure 215) appear in the second half of the 15th century, and, although their technique is essentially similar to that of the Syrian glassmakers, it is likely that they are of independent development. Little is known of the vessels made before this period, but it is evident from representations in pictures that they were mainly footed flasks and low beakers. The Venetians attributed the introduction of enamelling to a member of the glassmaking family of Barovier. The earliest pieces known, commencing with a goblet dated to 1465, certainly show no signs of outside influence. These, like most Venetian glass of the period, were inspired by the artistic ideals of the Italian Renaissance. The decorations represent triumphs, allegories of love, grotesques (fanciful combinations of human and animal forms), and so forth, with borders of dots of enamel laid on a ground of gold etched in scale pattern. Many of these pieces were of richly coloured glass, blue, green, or purple.

The Venetians were keenly aware of Roman achievements in glassmaking as in the other arts; they reproduced mosaic, millefiori, and aventurine glass, and glass resembling natural layered stones (*calcedonio*, sometimes miscalled *Schmelzglas*), and they even copied a Roman form of bowl that had vertical, external ribs. All these types of glass were Venetian specialities, and they were probably developed as a part of the extensive local bead industry.

Clear, colourless *cristallo* glass

The greatest achievement of Venice, however, and that upon which its great export trade came to be based, was the manufacture of clear, colourless glass, which was apparently exclusive to Italy during the Middle Ages (Figure 216). From its resemblance to natural crystal, this material was called *cristallo*, although in fact it often has a not unpleasing brownish or grayish cast. Made with soda, it was very ductile and cooled quickly. It therefore demanded of the workmen great speed and dexterity, and this, in turn, affected the nature of the glasses made. In the first half of the 16th century the Venetian glassblowers produced glasses of an austere simplicity. As the century proceeded (and more markedly still in the 17th century), however, there was a tendency to produce elaborate and fantastic forms. Enamelling on glass went out of fashion in Venice (except on pieces for export) in the first half of the 16th century. Its place was taken to some extent by the use of opaque white glass threads for decorative purposes (latticinio). This form of decoration became progressively more complex (Figure 217); opaque threads were embedded in a matrix of clear glass and then twisted into cables, which were themselves used to build up the wall of a vessel. The height of complexity was reached when a bulb of glass decorated with cables or threads running obliquely in one direction was blown inside a second bulb with threads twisted in the other direction. The composite globe thus formed was then worked into the desired form. This resulted in a vessel completely covered with a lacy white pattern (*vetro di trina*). Other methods of decoration at this time were mold blowing and dipping a vessel while hot into water or rolling it on a bed of glass fragments to produce a crackled surface (ice glass). *Cristallo* was also found suitable for engraving with a diamond point, a technique which produced spidery opaque lines that were especially suitable for delicate designs. The technique seems to have come into use about 1530.



Figure 215: Goblet, green glass enamelled and gilt, Venetian *c.* 1500. In the British Museum. Height 22.2 cm.
By courtesy of the trustees of the British Museum



Figure 216: Venetian glass ewer in the form of a nef ('ship'), attributed to Ermonia Vivarini, *c.* 1570. In the British Museum. Height 34.3 cm.
By courtesy of the trustees of the British Museum

Spread of Venetian glass-making

The glassworkers of the island of Murano were forbidden to leave Venice or to teach their secrets to outsiders, under dire penalties both to themselves and their families. Such was the demand for Venetian glass in the rest of Europe, however, and such was the desire of kings and nobles to control and reap the profits of its manufacture, that many Venetian workmen in the course of the 16th century were tempted to abscond to other countries, where they helped to set up glassworks. Furthermore, at Altare, near Genoa, existed a second great centre of glassmaking. Its glass was so like the Venetian in style and material that it is nowadays impossible to distinguish between the two. The



Figure 217: Tazza, colourless glass with opaque white striped decoration, Venetian, 16th century. In the Corning Museum of Glass, New York. Height 13.7 cm, diameter of rim 15.2 cm.
By courtesy of the Corning Museum of Glass, New York

glassworkers at Altare, moreover, were governed by no such laws as the Venetians; rather, they made it their policy to supply their men and teach their methods wherever there was a demand. Thus, the fugitive Venetians and the willing Altarists spread the Italian art of glass to the rest of Europe, and glasshouses were established in France, Spain, Portugal, Austria, and Germany, while in the North, Antwerp was a secondary source of diffusion.

Italian glassworkers ranged as far north as England, Denmark, and Sweden. Their labour was necessarily diluted by that of native workmen to whom they were often required to teach their methods. Variations in locally available raw materials modified the quality of the glass, and local taste influenced the form and ornamentation of the objects they produced. Nevertheless, in the late 16th and the 17th centuries an international style in glass developed, wholly Italian in origin and inspiration (*façon de Venise*).

Although there was everywhere a family likeness among glasses of the *façon de Venise*, certain countries developed types peculiar to themselves that are worthy of mention. Thus in Spain not only were fantastic and even bizarre shapes evolved in green glass, but in Barcelona a characteristic kind of enamelled decoration was developed, the peculiarities of which include a light-leaf-green colour and a constantly recurring lily-of-the-valley motif (late 15th–16th century). Elsewhere, at Hall, in the Tirol, a characteristic decoration with the diamond point, often supplemented by cold painting (*i.e.*, unfired oil—or other paint applied to a finished object), was favoured in alternating broad and narrow upright panels containing symmetrical scrollwork or coats of arms and other devices. Almost equally stiff and formal diamond-point work is to be seen on glasses probably made at the London glasshouse of Jacopo Verzelini (examples dated between 1577 and 1590). A more promising development of diamond-point engraving occurred in the Netherlands. There too the work of the 16th century was relatively formal and stiff, linear and clear, with simple hatching only. In the succeeding century, however, diamond-point engraving became initially more supple and pleasing, only to degenerate eventually into over-elaboration.

Diamond-point engraving in the Netherlands

Diamond-point engraving was practiced there widely by talented amateurs in the 17th century, among them Humanists such as Maria Tesselschade Roemers Visscher, her even more famous sister Anna Roemers Visscher and Anna Maria van Schurman. The latter two decorated their glasses with flowers and insects drawn with a gossamer touch, often accompanied by epigrams in Latin or Greek capitals scratched with severe precision or in the free scrolled style of the Italianate writing masters of the time. A similar calligraphy was practiced later in the century by the amateur Willem Jacobsz van Heemskerk, with notably beautiful results.

Engraving in the first half of the 17th century gradually abandoned linear clarity in favour of crosshatched chiaroscuro (shading) effects, the highlights formed by sometimes completely opaque spots. Many artists worked in this manner; two are worthy of special mention. One was an accomplished engraver signing "C.J.M.," whose earliest dated glass is of 1644; the other was Willem Mooleyser, of Rotterdam, who worked in the last two decades of the 17th century with a scribbled freedom and vigour that raised his work above the average. By the end of the century this type of diamond-point work was superseded in popularity by wheel engraving.

GERMANY

In Germany toward the end of the 17th century a reaction to Venetian glass styles seems to have set in. In that country there had been a continuous survival, probably from late Roman times, of a local type of green glass, a product of forest glasshouses made with potash obtained by burning forest vegetation and called therefore Waldglas ("forest glass"). From this material, often of great beauty of colour, were made shapes peculiar to Germany, notably a cylindrical beer glass studded with projecting bosses, or prunts (*Krautstrunk*, or "cabbage stalk"), and a wineglass (*Römer*) with cup-shaped or ovoid bowl set on a similarly prunted hollow stem (Figure 218). This became the classic

The Waldglas tradition

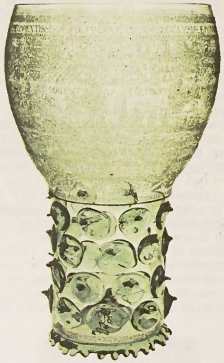


Figure 218: "The Mainz Dean and Chapter *Römer*," with diamond-point engraving of the city of Mainz; Netherlandish, 1617. In the Bayerisches Nationalmuseum, Munich. Height 32.5 cm.
By courtesy of Bayerisches Nationalmuseum, Munich

German shape of wineglass, which survived into the 18th century and, with modifications, to the present day. Apart from these indigenous forms, German glass in Venetian-type *cristallo* developed local characteristics of its own in the latter part of the 17th century.

In Nürnberg, for instance, the tall-stemmed Italianate goblet underwent a transformation into a severe glass with stem composed of no more than a baluster-shaped element and a bulb, which were joined together by a number of disk-shaped elements, or mereses, and attached to foot and bowl by the same means. Such goblets display some of the most accomplished glass engraving that has ever been practiced.

The leader and founder of the Nürnberg school of engravers was Georg Schwanhardt, a pupil of Caspar Lehmann. Lehmann had been gem cutter to the emperor Rudolf II in Prague and there had taken the decisive step of transferring the art of engraving from precious stones to glass. His first dated work is a beaker of 1605; in 1609 he obtained an exclusive privilege for engraving glass. Although he is the first great personality in glass engraving, he was not the first to practice the art in the German area.

On Lehmann's death in 1622 Schwanhardt inherited his patent and moved to his own native city, Nürnberg, where a whole school of glass engraving grew up around him and his family. Schwanhardt's work is characterized by delicate, tiny landscapes, often accompanied by bold formal scrollwork. His son Heinrich excelled in minute landscapes but also engraved inscriptions of fine calligraphic quality. Other notable Nürnberg engravers of the late 17th century were Paulus Eder; Hermann Schwinger, a master calligrapher; and H.W. Schmidt and G.F. Killinger, both notable for the delicacy with which they rendered landscapes. Somewhat similar work was done at Frankfurt am Main by members of the Hess family.

In Bohemia, after Lehmann's death, little engraving of high quality was done. Just before 1700, however, with the perfection of a massive, crystal-clear, potash-lime glass that allowed cuts of considerable depth, the engravers of the Bohemian–Silesian area came into prominence. The harnessing of the mountain streams in the Riesengebirge for water power enabled engravers (those of the Hirschberger Valley in particular) to practice relief engraving, which demands immense energy for grinding down the background of the design. Massive covered goblets were decorated with powerful acanthus scrolls in the contemporary baroque taste. Relief engraving (*Hochschnitt*) was only occasionally used by itself in the Bohemian–Silesian area in the 18th century; more often it was employed in conjunction with intaglio (*Tiefschnitt*). By the turn of the 18th century the engravers of this area—anonymous workmen regarded as artisans rather than as artists—had acquired great technical skill; this enabled them to adapt to glass all the changing fashions of the 18th century in the decorative arts. Glass engraving, often of fine quality, was also practiced in many parts of Germany—notably Thuringia, Saxony, and Brunswick—but the most significant work of the late 17th and early 18th centuries was that done in Brandenburg. There, the glassworks at Potsdam (moved to Zechlin in 1736) produced massive goblets and beakers that were engraved—usually to order for the court—in Berlin, where a water-powered engraving shop had been installed in 1687. Both relief and intaglio engraving were practiced, the latter being favoured. This workshop, indeed, produced perhaps the greatest of the German intaglio engravers, Gottfried Spiller, whose deep cutting on the thick Potsdam glass has seldom, if ever, been surpassed (Figure 219). A notable, if lesser, engraver from the same shop was Heinrich Jäger; and later, in the 1730s and 1740s, work of high quality was done by Elias Rosbach.

Bohemian relief and intaglio engraving

Another workshop of great significance was established toward the end of the 17th century at Kassel, in Hesse. There perhaps the greatest of all the relief engravers, Franz Gondelach, handled glass with a truly sculptural feeling.

In the second half of the 18th century, engraved glass declined in favour, although the technical skill required for its production never died out in the Bohemian–Silesian area. It experienced a great revival in the second quarter of the 19th century, when the taste of the newly prosperous bourgeoisie favoured elaborate decoration. The engraving of this period is often skillful in the extreme, although marred by excessive naturalism. Striking innovations of the period were the use of a casing (normally ruby red, blue, or opaque white) through which the design was cut down to the colourless glass. A yellow coating (the silver stain of the stained-glass artist) was often used in the same way. Notable engravers of this epoch were Dominik Bimann, August Böhm, A.H. Pfeiffler, and members of the Pelikan and Simm families.

German enamelled glass

Second in importance only to engraving as a method of decorating glass in Germany was enamelling. Germany had proved a profitable market for enamelled Venetian glass during the 16th century, and, in the latter part of that century, glass enamelling began to be practiced in the Germanic lands themselves, most notably in Bohemia. This enamelling, in bright opaque colours, was much favoured throughout the 17th century, chiefly on the cylindrical drinking glasses, often of great size, known as *Humpen* (Figure 220). The glass they were made of was often impure and of a greenish or yellowish cast, while the painting itself was the simplified repetitive work of artisans rather than of original artists. Nonetheless, the gaiety of colour of these glasses and a certain naiveté in their painting give them an authentic unsophisticated charm. The most favoured types of decoration include a representation of the imperial double-headed eagle (*Reichsadlerhumpen*); representations of the emperor with his seven electors, either seated or mounted on horseback (*Kurfürstenhumpen*); subjects from the Old and New Testaments; and allegorical themes such as the Eight Virtues and the Ages of Man. These were painted between borders of multicoloured or white dots or intersecting ellipses, often on a gold ground. This general style continued into the 18th century; but in the course of that century the levels of artistic and technical competence sank, and the tumblers and spirit bottles, which were the main types produced, can be regarded only as objects of peasant art.



Figure 219: Beaker and cover, unpolished intaglio engraving with relief-cut laurel frieze by Gottfried Spiller, *c.* 1700. In the Kunstmuseum Düsseldorf, Germany. Height 27 cm.
By courtesy of Kunstmuseum Dusseldorf, Germany

SCALA—Art Resource/EB Inc



Figure 220: *Humpen* (enamelled drinking vessels), German, 17th century. (Left) Tankard decorated with a representation of the Trinity. Height 30 cm. (Right) *Reichsadlerhumpen*, decorated with the imperial double-headed eagle. In the Germanisches Nationalmuseum, Nürnberg. Height 28 cm.



Figure 221: Beaker painted in two tones of black enamel (*Schwarzlotmalerei*) by Johann Schaper. Nürnberg, 1664. In the Museum für Kunst und Gewerbe, Hamburg. Height 8.9 cm.
By courtesy of Museum fur Kunst und Gewerbe, Hamburg

The *Schwarzlotmalerei* technique

A far more sophisticated type of enamel painting was carried on during the third quarter of the 17th century at Nürnberg. There, painting in black or sepia (*Schwarzlotmalerei*)—a technique borrowed from the stained-glass artist—was used to decorate the small cylindrical beakers (often resting on three hollow ball feet), which were a locally favoured shape. Other colours, notably red used in touches with the black, were occasionally employed. The greatest and most original artist of this school was Johann Schaper, who painted delicate architectural and landscape compositions in which a fine point was used to etch in details (Figure 221). The best of Schaper's followers were J.L. Faber, Hermann Benckerlt, Johann Keyll, and Abraham Helmhack, but none of them equalled him in artistic competence. Comparable work appears to have been done, although on a more restricted scale, in the Rhineland, notably by Johann Anton Carli of Andernach. At the beginning of the 18th century *Schwarzlot* painting, often with touches of gold, was practiced in Bohemia and Silesia and reflected the changing fashions in the decorative arts. Daniel Preissler and his son Ignaz are known to have done this work.

In the first half of the 19th century the decorators of vessel glass once again borrowed from the stained-glass artist. Samuel Mohn, his son Gottlob Samuel Mohn, and Anton Kothgasser painted the beakers typical of this "Biedermeier" period in transparent enamels and yellow stain.

By courtesy of Kestner Museum, Hannover, Germany

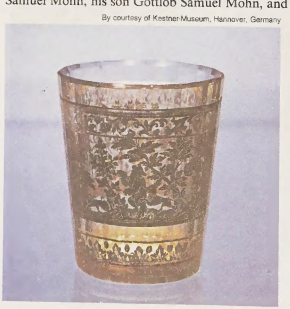


Figure 222: *Zwischengoldglas* ("gold sandwich glass"), double-walled beaker decorated with a bear hunt, Bohemian, *c.* 1730. In the Kestner-Museum, Hannover, Germany. Height 8.9 cm.

The *Zwischengoldgläser* technique

A technique peculiar to Bohemia in the 18th century was that of the "gold sandwich glasses" (*Zwischengoldgläser*). These were beakers or less often goblets made of two layers of glass, exactly fitting one over the other, between which was sandwiched a gold leaf previously etched with a steel point to the desired design (Figure 222). The earliest work in this technique was anonymous, but late in the century J.J. Mildner employed it with notable success, making gift tumblers decorated with medallions of etched gold or silver leaf (often backed with red pigment) and sometimes also engraved on the wheel or with the diamond point.

ENGLAND

Glass was certainly made in England during the later Middle Ages, but most of it was used for church windows (see *Stained glass*). The vessel glass of the period has not been much studied and is only imperfectly understood. Only by the second half of the 16th century does the picture become clearer. Two lines of development may be traced in this period. One is the glass of German waldglas type, made in the woods that supplied the furnaces with fuel and a source of potash. These glasses were made by workers whose traditions were those of Lorraine and northern France. Much of their production was of window glass, but they also made vessels in a modest variety of shapes and modes of decoration. Chief among them was a tumbler-like drinking glass with a low, double foot-rim produced by pushing in the bottom of the bulb from which the glass was made; this might be decorated either by mold-blown diaper (overall repeat) patterns, by swirled ribbing imparted by mold blowing and subsequent twisting, or by a zone of trailed threading below the rim. Applied notched ribbons or small circular motifs also were used. Small bottles of mold-blown hexagonal section or of flattened ovate form with diagonal ribbing also were made. The second line of development was that of the international Venetian style brought by immigrant Italians; this, however, in time acquired an English idiom. The work was done mainly in London.

17th-century English glass

In the 17th century these two traditions were welded into one, spurred by the proclamation of 1615 that forbade the use of wood in glass furnaces, as well as in certain other industries, in an effort to prevent the deforestation of the country. Thereafter, with coal as the sole means of fusing glass, glassworks tended to be located where coal deposits (and the frequently concomitant fire clays for making glass pots) were abundant. Since such areas for the most part were those that have been continuously occupied by industry (*e.g.*, the Stourbridge area and Tyneside), exploration of the early glass factory sites has seldom been practicable. Little, therefore, is known of provincial glass-making in England in the 17th century, but it is clear that Venetian influences gradually replaced the earlier waldglas tradition, which had depended on supplies of wood. Some idea of the new style may be gained from the fragments of glasses often excavated in London and other cities. It is frequently difficult to distinguish between an English glass and an imported European one, although a certain coarseness may be taken as symptomatic of English make.

During the first half of the 17th century, glassmaking was among the English industries for which monopoly rights were granted by the crown; the greatest of a series of monopoly holders was Sir Robert Mansell, who effectively controlled the industry from 1623 until his death in 1656. After the Restoration, although some monopolies were granted for certain categories of glasswares, an increasingly important role in the English industry was played by the Worshipful Company of Glass Sellers (reincorporated in 1664), which was able to keep closely in touch with the needs of the English market. Its members seem to have laid stress on simplicity of shape and durability of material, as appears from the correspondence of one of them, John Greene, with his suppliers in Venice. Dissatisfied with the quality of glass supplied to them and no doubt also anxious to make England independent of foreign sources of both finished glass and raw materials, they commissioned George Ravenscroft to make experiments with

native materials in the hope of evolving a more solid glass than the Venetian and one that more closely resembled rock crystal.

Discovery of lead glass

Ravenscroft was completely successful; his crucial discovery was the value of adding lead oxide. His "glass of lead," evolved about 1675, was perfected toward the end of the century and set a standard for the rest of Europe. It was solid and heavy and more durable than the Venetian-type glass, which it progressively displaced. It was also characterized by brilliance and dark shadow paradoxically combined. It was slower to work than the Venetian glass and gradually the Venetian idioms were dropped from English glassmaking in favour of a genuine native style. This style is best exhibited in the drinking glasses that, by the end of the 17th century and the beginning of the 18th, constituted the chief glory of the English industry. These often massive baluster-stem glasses were composed of a usually funnel-shaped bowl and a stem compiled of any of a large variety of pear-shaped and bulbous knops (ornamental knobs). In their simplicity and the harmony of their proportions they rank among the classics of the Queen Anne style.

Toward the middle of the 18th century, taste in the arts generally inclined to lighter forms, and in glass this tendency was given additional impetus by an excise (1745–46) levied on glass by weight. Drinking glasses became slighter, the bowls smaller, and the stems taller and more slender. The loss in architectonic values was often offset by extraneous decoration. At first this tended to be concentrated in the stem. Bubbles of air had sometimes previously been enclosed in a knop forming part of the stem of a wineglass, and these bubbles were now drawn out and twisted so that they formed a cable of air ribbons inside a cylindrical stem. Stems of this type were popular about the middle of the century. Just before 1750 a stem decorated with threads of opaque white glass instead of air twists came into favour. These stems were made by much the same techniques as the Venetian latticinio glass. They remained in fashion until about the time of the second Glass Excise Act in 1777, which imposed a tax on the opaque white "enamel" glass, previously exempt.

These forms of ornament had been restricted to the stems of glasses, but other methods of decoration were simultaneously evolved to embellish the whole glass. First of these was engraving, which had been sporadically practiced in England as early as the end of the 17th century. This work and the inscriptions, coats of arms, and arabesque borders in German style that were engraved during the first 20 years or so of the 18th century were undoubtedly the work of immigrant (probably German) artisans. By 1735, however, at least one English engraver was capable of executing such commissions and from about this time engraving on glass began to take on a more English character. An artless use of floral motifs, chinoiseries (Chinese themes), and scenes from country life is typical of the engraving of the third quarter of the 18th century, as were the frequent representations on glasses of Jacobite themes—portraits of the Old and Young Pretenders (James III and Charles Edward), the rose with buds, the honeysuckle, and the other flowers used in the symbology of the Stuart cause, together with the mottoes of such "loyal" societies as the Cycle Club.

Development of engraving and enamelling in England

Engraving never reached great heights in England, but English glasses were in demand by engravers in Europe, particularly in the Netherlands, where the work of at least one notable artist—Jacob Sang, of Amsterdam—was almost exclusively done on imported English drinking glasses. English lead glass also seems to have been particularly favoured by the Dutch diamond-point engravers, whose work in this period was executed almost exclusively in stipple (*i.e.*, dotted engraving). The chief masters of this delicate art, in which the design seems no more than a bloom on the surface of the glass, were Frans Greenwood of Dordrecht, the originator of the style (Figure 223), and David Wolff of The Hague, whose work, if uninspired, is of high technical accomplishment.

Enamelling, the second decorative technique of foreign inspiration, began to be used on English glass in the mid-18th century. It embellished opaque white glass in imitation of china—a type of work usually associated with the name of Michael Edkins, a Bristol artist, but in fact done in many parts of the country. Perhaps the most original work in this medium was done on clear glass by members of the Beilby family of Newcastle upon Tyne during the 1760s and 1770s. Their rendering in usually blue-toned white enamel of ruins, trophies of arms, and rural pastimes, often framed in scrollwork of the utmost delicacy, is one of the best things in English Rococo glass. Gilding was also used at this time to decorate glasses, usually with simple designs of vines and grapes.



Figure 223: Glass goblet with diamond-point stipple engraving; signed "F. Greenwood fecit 1764." Holland. In the Museum für Kunst und Gewerbe, Hamburg. Height 28 cm.
By courtesy of Museum für Kunst und Gewerbe, Hamburg

These ornamental techniques, however, were of ephemeral growth in England. Far more significant than any of them, because more firmly rooted in the very nature of English glass, was the art of cutting. Although literary references to cut glass occur before 1720, the earliest known pieces can hardly be dated much before 1725. On them the cutting is mainly confined to brims and feet, which are scalloped or notched; or, on wineglasses to the thicker parts of the glass, such as the stem, which might be fluted or cut in an allover pattern of flat diamonds. Throughout the period from about 1745 to 1770, shallow cutting was the norm. Diamonds, hexagons, flutes, and scale pattern were combined with segmental lunate cuts (produced by holding the glass at an angle to the cutting wheel) and with triangular and diamond motifs in very low relief. All of these elements could be combined to produce designs of great complexity and richness. This period marked the golden age of English cutting.

Importance of cutting as a decorative technique

About 1770 a plainer style, employing mainly flutes, responded to the rising Neoclassical fashion in the other arts. The flutes were sometimes combined with diamonds in relief. When further taxes were imposed on glass in 1777 and 1781 and when in 1780 trade between England and Ireland was freed, it was this relief-diamond style that was taken up in Ireland by the glasshouses founded there. The Irish glassworkers could afford to be more lavish with their material and on this thicker glass increasingly deeply cut diamonds and other relief motifs could be produced. About the turn of the century the diamonds began to be reduced in size and to be incorporated into a diaper pattern covering whole areas, often alternating with fields of larger truncated diamonds, the surfaces of which were themselves diversified with cut crosshatching. Such designs were often combined with deeply cut horizontal grooves. These styles, which were subsequently followed

in England as well as in Ireland, finally led to a complete breaking up of the face of the glass into points and ridges, with increased prismatic effect but with a disastrous loss of surface quality, which is one of the peculiar beauties of glass. The prismatic brilliance was enhanced by the progressively greater purity and whiteness of the glass made during the second quarter of the 19th century. The temptation to cut ever more deeply and with greater complexity finally seduced the glassmakers into producing the "prickly monstrosities" of the Great Exhibition of 1851.

Throughout the 18th century there had been great admiration in Europe for English lead "crystal," and in the second half of it some of the European glasshouses were using lead oxide and had contrived to produce a comparable material. English cut glass was admired and exported, and the styles of cutting of the late 18th and early 19th centuries were much imitated abroad. (R.J.Ch.)



Figure 224: Sugar bowl with cover, pattern molded, attributed to the glassworks of Henry William Stiegel, Manheim, Pennsylvania, c. 1765–74. In the Corning Museum of Glass, New York. Height 15.6 cm.
By courtesy of The Corning Museum of Glass, New York

UNITED STATES

Glassmaking was apparently the first industry to be transplanted from Europe in the wake of the Spanish conquerors. As early as 1535 glass was being made at Puebla in Mexico, and in 1592 a glasshouse was located in the territory of the Río de la Plata in the town of Córdoba del Tucumán, Argentina. Broken glass, undoubtedly of European origin, was remelted at Córdoba and fashioned into various objects including thick, semitransparent flat glass.

The London Company of Virginia set up a glasshouse in Jamestown in 1608 for the manufacture of "glasses" and beads. A "tryal of glasse" was sent off to England before the winter of 1609, the "starving time" during which 440 of the colony's 500 inhabitants died. In 1621 the company tried again and, although the second attempt was more carefully planned, it too failed. Excavation of the site has revealed that glass was melted in considerable quantities though no evidence of glass bead manufacture has been found.

South Jersey-type glass. For more than a century after Jamestown, there was little American glass. The earliest successful glasshouse was begun in 1739 by Caspar Wistar in Salem County, New Jersey. The fact that his works produced only humble utilitarian vessels and windowpanes saved him from extermination by the "lords of trade." Wistar died in 1752, after which the factory was operated by his son Richard. It was offered for sale in 1780. Although few, if any, objects exist that can be assigned to the Wistar Glass Works with certainty, it is important as the cradle of the American glass known today as South Jersey type. That glass is the work of individual glassblowers using ordinary bottle or window glass to make objects of their own design. Applied glass and, occasionally, pattern molding were the only feasible means of decoration, and the resultant loopings and threadings are typical of European traditions. One decorative device, the lily pad, is of particular importance, as no European prototype is known. A hot mass of glass applied to the base of the bowl is pulled up around the sides in a series of projections in which the bowl appears to rest.

Wistar's glasshouse

Stiegel glass

The second great name in early American glass is Henry William Stiegel. Like Caspar Wistar, Stiegel at first was concerned with the manufacture of bottles and windowpanes, which he began in 1763 at his iron forge in Lancaster County, Pennsylvania, and continued in his new glasshouse at Manheim, also in Lancaster County, sometime after 1765. Encouraged by the patriotic adoption of the non-importation agreement, he ventured into the table-glass business, running many advertisements in which he favourably compared his wares with English imports. Later called the American Flint Glass Works, it failed in 1774 after adverse economic conditions, caused by both the approaching war and the colonial preference for imported tablewares.

Few pieces can be attributed with confidence to the Stiegel factories, and, like that of Wistar, his name survives as the founder of a tradition (Figure 224). Stiegel-type glass is characterized by the use of clear and artificially coloured glasses; by extrinsic decoration such as engraving, enamelling, and pattern molding; and, in general, by two distinct styles, one employing English and the other German techniques and decorative devices. Certain mold-blown patterns, such as the diamond daisy and daisy in hexagon, are believed to have been originated at the Stiegel houses, no European prototypes having been identified.

Post-Revolutionary glassworks. Before the turn of the century, several other glassworks were founded, but few survived the Revolution. These houses were devoted largely to the manufacture of bottles and window glasses and, with the notable exception of the New Bremen Glassmanufactory, most of the offhand (*i.e.*, shaped by hand) pieces that can be tentatively assigned to them are of the South Jersey tradition. Three of these enterprises are of particular importance. First, the New Bremen (Maryland) Glassmanufactory, founded by John Frederick Amelung and Company, is of special interest as many of its presentation pieces are both signed and dated as well as being among the finest produced in the United States before

By courtesy of the Metropolitan Museum of Art, New York, Rogers Fund, 1928



Figure 225: Bremen Pokal, presentation piece engraved at the glassworks near Frederick, Maryland; inscribed "Old Bremen Success and the New Progress and New Bremen Glassmanufactory—1788—North America, State of Maryland." In the Metropolitan Museum of Art, New York. Height with cover 28.6 cm.

1800. Originally from Bremen, Germany, Amelung was persuaded to go to America for the express purpose of founding what he believed to be a much-needed industry. By 1795 his works offered green and white hollow ware for sale; by 1795 the glassworks themselves were offered for sale. One of the most famous pieces in the history of American glass is the Bremen Pokal (the German word for goblet), blown and engraved in 1788 (Figure 225) and sent back to Amelung's financiers in Bremen, probably the only return they ever received on their investment.

The second factory of importance, later known as the Olive Glass Works, Gloucester County, New Jersey, was completed in 1781 by former employees of the Wistar Glass Works, the Stanger brothers. In addition to the many fine South Jersey pieces attributed to this house, it is of interest because of its long history, eventually becoming part of the Owens Bottle Company, a forerunner of Owens-Illinois, Inc.

The Pitkin glassworks

The third notable venture begun before 1800 is the well-known works associated with the name Pitkin. Erected at East Hartford, Connecticut, near the Connecticut River in 1783, it was intended for the manufacture of window glass, but in 1788 it was converted to the manufacture of bottles and flasks. The factory thrived until 1830 and is best known for the half-post (*i.e.*, dipped twice up to the neck) ribbed flasks in natural browns, ambers, and greens. Today the word Pitkin denotes a type of flask and not a specific glassworks.

After the War of 1812. The few houses that survived the 1790s and the depression after the War of 1812 had multiplied to more than 90 by 1830. For convenience, the glassworks are divided into three geographical groups: New England, the Middle Atlantic, and the Midwest. Until that time, they had produced little more than simple imitations of European glasses, at best interesting and often very handsome combinations of various decorative devices and traditions. The big change occurred between 1830 and 1840 with the production of fine lead glass, the use of the full-size incised mold, and, finally, the pressing machine.

Bakewell's fine glass

The glasshouse known as Bakewell's was synonymous with the finest achievements of the revived industry. Originally established in 1808 in Pittsburgh, the first city to use coal for fuel in glassmaking, the company survived under several different firms until 1882. Glass cutting, introduced to Pittsburgh by William Peter Eichbaum, glass cutter to Louis XVI, was an important part of Bakewell's operation. In addition to being the first American company to supply the White House, serving President James Monroe in 1817, Bakewell's produced such specialties as lead-glass tumblers with "sulphides" (cameo insertions of white fireproof material in an envelope of glass) in the bases portraying the Marquis de Lafayette, Andrew Jackson, New York governor George Clinton, Benjamin Franklin, and George Washington. The company also held the first patent on mechanical pressing, granted in 1825 for a device to make knobs.

Fine lead glass in the New England area was first successfully made in the South Boston works of the Boston Crown Glass Company. Thomas Cains was making flint glass there in 1813. He left the firm in 1824 to found the Phoenix Glass Works in South Boston, which survived until 1870. One particular device usually associated with the Boston manufactories of this period is the guilloche, or chain, employed in the decoration of a large variety of tableware.

The New England Glass Company, founded in 1818 in Cambridge, Massachusetts, maintained the same high standards as Bakewell's, even to the point of making glass for President Monroe. This factory held the second patent on a device for mechanical pressing, granted in 1826, and produced quantities of pressed glass of all types before it was moved to Toledo, Ohio, in 1888. The New England Glass Company was also famous for its very fine freeblown and engraved glass. In addition, vessels were made there in the so-called blown three-mold technique, in which decorative designs adapted from cut-glass patterns of the period were impressed in the glass by blowing in molds hinged in two, three, or more sections. More than 400 different molds have been determined and grouped according to pattern under three primary headings: geometric, arch, and Baroque. By 1830 this type of production was being replaced by the much more efficient pressing machine.

The three-mold technique

Deming Jarves, one of the founders of the New England Glass Company, founded the Boston and Sandwich Glass Company in 1825. Because of his *Reminiscences of Glassmaking*, extensive advertisements, and thorough excavations of the factory site in Sandwich, Massachusetts, more is known about this particular factory than any other of the period. Consequently, "Sandwich" has become a generic term for pressed glass even though many other factories used identical machinery and, in some cases, identical molds. Jarves's first patent on a pressing device, the fifth to be granted, was received in 1828 after the Boston mold maker Hiram Dillaway entered his employ. Jarves founded the Mount Washington Glass Works in 1837 in New Bedford, Massachusetts, and the Cape Cod Glass Works in 1857.

Among the outstanding makers of fine lead glass in the middle Atlantic states were the Brooklyn Flint Glass Works of John L. Gilliland and Company and the Dorflinger Glass Works. Gilliland, a partner in the Bloomingdale Flint Glass Works, sold out in 1823 and founded his own works in Brooklyn, New York. In 1864 two members of the Houghton family acquired controlling interest, and in 1868 the works was moved by barge to Corning, New York, to form part of the now famous Corning Glass Works.

Historical flasks. Perhaps the most fascinating aspect of American glass is a series of pictorially molded bottles known as historical flasks, produced between 1815 and 1870. Some three hundred ninety-eight different surviving examples have been divided into the following groups: (1) Masonic; (2) emblems and designs related to economic life; (3) portraits of national heroes and designs associated with them and their deeds; and (4) portraits of presidential candidates, emblems and slogans of political campaigns. In the second group are a number of interesting designs encouraging the United States system of better internal transportation and high protective tariffs. Among the 16 celebrities portrayed in the third and fourth groups are Jenny Lind, the Swedish singer; Lajos Kossuth, the Hungarian patriot (Figure 226); Marquis de Lafayette, the French hero of the American Revolution; and the notorious Thomas W. Dyott, a patent-medicine vendor and bottle manufacturer. These containers were used also as propaganda during political campaigns. William Henry Harrison is pictured in this connection with other impedimenta relative to the "Log Cabin and Hard Cider" campaign of 1840.

The "lacy period" of pressed glass

The first 25 years of pressed glass, 1825 to 1850, are referred to by collectors as the "lacy period." A mile-



Figure 226: "Portrait of Kossuth" flask, Bridgeton, New Jersey, 1840–55. In the Corning Museum of Glass, New York. Height 17.5 cm.
By courtesy of the Corning Museum of Glass, New York

stone within this brief span occurred in 1830 with the development of the cap ring, a device that ensured uniform thickness at the edge of each piece regardless of the amount of glass forced into the mold. Before this date most impressed designs were inspired by Anglo-Irish cut glass, often coupled with popular American devices such as a sheaf of wheat. Between 1830 and 1840 the objects were thinner and more lavishly decorated, often including elaborate motifs based on the classic and Gothic revivals. Because of the unpleasant surface left by the mold and in an effort to imitate the brilliance of cut glass, unstippled areas were filled in with overall lacelike patterns; hence the term "lacy." About 1840 economic conditions forced glassmakers to revert to cheaper molds and simpler geometric forms and to abandon the stippled patterns.

During this period the mechanical press became firmly established, and by mid-century glassmaking had become one of the United States' new mass-production industries. (T.S.Bu.)

Mid-19th to 20th century

The modern history of glass can be said to begin in the middle of the 19th century with the great exhibitions and with the new self-consciousness in the decorative arts that they expressed. Glassware was being publicly discussed in art journals and collected in museums, and this new spirit of awareness led to a greatly increased exchange of ideas among the leading glass centres and to the borrowing of ideas from the past.

In some degree the established glass-producing centres were still concerned in the modern period with the styles of glassware for which they had achieved an earlier reputation. The English glasshouses continued their production of deeply cut crystal; engraved glass and to a lesser extent coloured and painted glass were given the greatest attention in central Europe; the Venetian glasshouses at Murano were the leading exponents of furnace-manipulated glass. But alongside these traditional methods of using and decorating glassware can be discerned the development of a renewed interest in the beauty of the material itself. Expressed in various ways, in the use of thick masses and in internal figuring and patterning, this interest has been the keynote of the most significant modern contributions to the art of glass.

Pressed glassware, which had been first made with great promise in the first half of the 19th century, was being widely made in the middle of the century, and later, as a cheap imitation of cut crystal. The decorative possibilities of the process continued, however, to be exploited in a variety of popular wares; and in the 20th century a series of new simple forms of pressed glassware appeared that had been expressly designed in relation to the characteristics of its manufacture.

GREAT BRITAIN

The Great Exhibition of 1851 was the culmination of a period of intense activity in the British glasshouses. The excise duty on glass had been removed in 1845, and the British glassmakers were determined not only to excel in their traditional deeply cut crystal but also to rival the Bohemians and the French in coloured, layered, and enamel-painted wares. Probably the most enterprising of the English glassmakers of the period was Benjamin Richardson, of Wordsley near Stourbridge; surviving pieces of this period from the Richardson firm include some admirable painted and engraved pieces as well as crystal wares deeply cut in bold patterns.

Probably in reaction against the banality of pressed-glass imitations of cutting, the most sophisticated work in crystal during the later 1850s through the 1870s was decorated by engraving, often carried out by immigrant Bohemian craftsmen.

The Venetian style of furnace-manipulated glass was also exerting a strong influence. It can be seen, for instance, in the development of the elaborate Victorian centrepieces in the 1860s and 1870s. In some degree the Venetian style was also an influence, alongside that of the Far East, in the fashioning of the fancy wares that were made in Great Britain—as it was in the United States and elsewhere—during the 1880s and 1890s. These wares were often given specific trade names and were mostly made in the English Midlands by firms such as Thomas Webb & Sons of Stourbridge and John Walsh Walsh of Birmingham.

A striking form of mid-Victorian virtuosity was the cameo glass produced by Stourbridge glassworkers. This work, inspired by the Portland vase, required a lengthy process of etching and carving, normally through an opaque-white-glass layer to leave a white carved design in relief on a dark-coloured glass body. The first important pieces, such as the "Pegasus vase" (Figure 227), were produced in the

By courtesy of the Smithsonian Institution, Washington, D.C.



Figure 227: The "Pegasus vase," carved in cameo relief by John Northwood of Stourbridge, England. In the Smithsonian Institution, Washington, D.C. Height 54.6 cm.

1870s by John Northwood, and in the later part of the century the most distinguished cameo work was carried out by George Woodall.

The influence of the Arts and Crafts Movement was toward the use of plastic forms and furnace decoration, which the English art critic John Ruskin had advocated in *The Stones of Venice*. In 1859 Philip Webb designed for William Morris some simply formed tableware that was made at the London glassworks of James Powell & Sons. From about 1880 this glassworks was under the control of Harry J. Powell who, working until World War I, developed a simple, dignified style of handmade blown glass, which was subsequently continued in designs by Barnaby Powell, James Hogan, and others.

Influence of the Arts and Crafts Movement

During the 1930s and after World War II other firms produced work in which a restrained and distinctively modern approach was made to the cutting of faultless crystal glass. Notable designs were produced by Keith Murray for Stevens & Williams shortly before World War II and by David Queensberry (12th marquess of Queensberry) for Webb Corbett in the 1960s. Among the more distinguished glass engraving may be mentioned the diamond-point fantasies of Laurence Whistler and the work of John Hutton, made by a movable wheel held in the hand, such as his great screen in the new Coventry Cathedral. The appearance of new factories in the 1960s, concerned primarily with form and colour, widened the scope of British glass design; and at this time the glass-teaching schools

were especially significant as centres for original work by individual artists.

Importance of pressed glass

UNITED STATES

By the middle of the 19th century, American pressed glass was already a disturbing influence on the design of the finer wares. Its decoration was by that time mostly designed in imitation of cut glass, and the process of fire polishing was being used to give a surface almost as smooth as that of blown glass. During the succeeding decades pressed-glass designs became increasingly complicated. This tendency was accentuated in the soda-lime glass that William Leighton began to use for pressed work at Wheeling, West Virginia, in the 1860s, and that was later widely used in the western glasshouses for the cheapest coloured wares.

In general the finer wares of the early part of the period were similar to those of the Biedermeier and later styles of Europe. The New England Glass Company at Cambridge, Massachusetts, was employing many European craftsmen and was producing a wide variety of richly decorated layered and engraved wares. At the Boston and Sandwich Glass Company layered glass was extensively used for large kerosene lamps. The effect of the competition of pressed glass on cut-crystal work can be seen in the appearance of fine-line cuttings, and, during the period up to the Philadelphia Centennial Exposition of 1876, the most significant crystal work was decorated by engraving. Louis Vaupel and Henry S. Fillebrown were two notable engravers employed by the New England Glass Company from 1856 and 1860, respectively.

At the time of the Centennial Exposition, cut-crystal work began to revive, and by 1880 a considerable boom in its production had developed—a boom that was to continue throughout the 1880s and 1890s. New industrial methods contributed to the production of crystal glass of flawless quality and to its deep cutting with mathematical accuracy in elaborate designs. Among many others, a noteworthy producer of this type of glass in the 1890s and later was the Libbey Glass Company, the successor to the New England Glass Company. Later, in the early years of the 20th century, intaglio cutting in crystal became popular, and work in this expensive process was carried out in a number of cut-glass factories such as the T.G. Hawkes Glass Company at Corning, New York.

As in Great Britain and elsewhere, a great amount of glass was made in fancy forms and colours in the 1880s and 1890s. Although undisciplined and often tasteless, such glass nevertheless preserves perhaps more than any other the flavour of the period. These wares, often bearing specific names such as Pomona, Burmese, and Peachblow, were made by such firms as the New England Glass Company, the Mount Washington Glass Company at New Bedford, and the Hobbs, Brockunier Company at Wheeling, West Virginia.

Although belonging essentially to the category of the fancy glasses, the Favrile glass of Louis Comfort Tiffany (Figure 228) represented an altogether higher level of achievement both in its shapes and in the colouring and figuring of the glass. It was first shown to the public in 1893, and in pieces that were produced a few years later Tiffany achieved an outstanding expression in glassware of the Art Nouveau style. Much of his work was in a heavily lustred glass that was considerably admired abroad, especially in central Europe where it created a new fashion.

Tiffany's Favrile glass

By courtesy of The Metropolitan Museum of Art, New York, gift of H.O. Havemeyer, 1896



Figure 228: Peacock vase, Favrile glass by Louis Comfort Tiffany. In The Metropolitan Museum of Art, New York. 35.9 × 29.2 cm.

By courtesy of Steuben Glass



Figure 229: "Moby Dick," Steuben crystal sculpture designed by Donald Pollard and Sidney Waugh, first piece made in 1959. Length 28.6 cm.

From the period of World War I onward, new forms of pressed glassware appeared in simple, satisfying designs appropriate to their purpose and the process of manufacture, such as the Pyrex ovenware shapes of the Corning Glass Works. The Steuben Glass Company of Corning was known for fancy glasses designed by Frederick Carder, until in 1933 the company was given a change of direction by Arthur Amory Houghton, Jr., who, with the help of John Monteith Gates and the sculptor and designer Sidney Waugh, aimed to produce glass with engraved decoration that would rank as fine art (Figure 229). Other noteworthy modern American work included simple designs in blown glass by the Blenko Glass Company of Milton, West Virginia, and enamel patterned bowls by the independent artist Maurice Heaton. The appearance in the United States of studio blown glass, produced by individual artists, was a development of international significance. It was initiated in the 1960s notably by Harvey Littleton and Dominick Labino and included work such as that produced personally by Joel P. Myers at the Blenko Glass Company.

CZECHOSLOVAKIA, AUSTRIA, AND GERMANY

In the middle of the 19th century the glasshouses of central Europe were producing a great variety of the layered and coloured wares that had become particularly associated with Bohemia in the preceding Biedermeier period (Figure 230). They were also producing a great amount of cut crystal glass in the deeply cut English style, and indeed work of this nature continued with little change throughout the modern period.

Revival of engraving

A revival of the indigenous art of engraving was initiated



Figure 230: Bohemian layered-glass vase, painted and gilt by Wilhelm Hoffmann, Prague and Vienna, *c.* 1850–60. In the Victoria and Albert Museum, London. Height 42 cm.
By courtesy of the Victoria and Albert Museum, London

by Ludwig Lobmeyr, who from 1864 was in control of the Viennese firm of J. and L. Lobmeyr. His first opportunity came at the Paris International Exhibition of 1867, and his reputation was firmly established at the Vienna International Exhibition of 1873. He commissioned designs for his glasses from the leading Viennese architects and painters of the time, and his work was carried out by the finest craftsmen in Bohemia and Austria.

The Art Nouveau style, which went under the name of Jugendstil in central Europe, made a deep impression on central European glassware. The work made around the turn of the century abounds in slender shapes and flowing organic motifs. Glasses designed by Karl Köpping in Berlin, with long, waving stems and tulip-like bowls, were perhaps the extreme instance of Art Nouveau style applied to glassware. In 1897 an exhibition of glass by Tiffany was shown at several of the museums in the area. Not only the forms of the Tiffany glasses but also their figured and heavily lustred material attracted great interest. Several factories started making a similar heavily lustred glass, including the firm of J. Lötz' Witwe of Klášterský Mlýn (Klostermühle), which won a *grand prix* at the Paris Exhibition of 1900 with this type of glassware.

From around 1900 onward a movement toward a modern purist approach to glass was largely fostered by the work of designers connected with the Vienna Kunstgewerbeschule (School of Industrial Art). Men such as Kolo Moser and Josef Hoffmann, who were also closely associated with the Vienna Werkstätte (Workshop), were designing glasses in simple rational forms. Much initiative in this movement was shown by the firms of E. Bakalowits Söhne of Vienna and J. Lötz' Witwe. The Czech architect Jan Kotěra was influential in the modern design of glass, and in the early years of World War I the Czech Artěl organization of artists and architects was concerned with the design of glass in a forward-looking Cubist manner.

Post-World War I glass

After World War I the outstanding figure in Czech glass art was Josef Drahoňovský, who was professor at the Prague School of Industrial Art. He was essentially a sculptor, and most of his glass designs were for sumptuously engraved glass of a monumental quality. His colleague in Prague, Jaroslav Horejc, designed for engraved work of a broadly similar character, some of it for the Lobmeyr firm of Vienna. The decades after World War II saw considerable activity in glass design. Notable artists in the 1960s were Stanislav Libenský, René Roubíček, Pavel Hlava, and Václav Cigler.

In Austria after World War I the Lobmeyr firm under the control of Stefan Rath produced many engraved and relief-carved pieces designed by artists such as Ena Rottenberg, Lotte Fink, and Vally Wieselthier. Lobmeyr also produced some of the best designs of Michael Powolny, who had his own workshop and had designed for the firm of J. Lötz' Witwe.

In Germany the outstanding engraver and glass carver of the period after World War I was Wilhelm von Eiff, a professor at the Stuttgart Kunstgewerbeschule. Bruno Mauder of the glass-teaching school at Zwiesel in Bavaria advocated the use of natural and appropriate glass forms. Some fine tablewares were produced, especially after mid-century, by designers such as Wilhelm Wagenfeld, Richard Süssmuth and Heinrich Löffelhardt. An interesting development was that of the Rosenthal firm, which used the Finnish designer Tapio Wirkkala and the Dane Bjørn Wiinblad to effect in each case matching glass and porcelain suites of the firm's own manufacture.

FRANCE

In France, as in central Europe and in England, the production of fine glassware in the middle of the 19th century was mainly divided between cut crystal and coloured wares. The "opalines," the semi-opaque white and coloured wares, often with elaborately painted and gilt decoration, were especially popular; and it was during these years that the French paperweights, containing coloured patterns, became internationally known and admired. The larger factories, particularly Baccarat and Saint-Louis, continued to participate in the international fashions of the rest of the century and beyond. But in France inventive genius manifested itself mainly in the work of individual artists and thereby a new spirit was introduced into the modern conception of glass.

In the late 1860s and 1870s three individual artists were experimenting in glasswork, and all of them were represented in the International Exhibition of 1878 in Paris. The first was Joseph Brocard, who was studying the enamelling of glass and whose main ambition was to reproduce medieval Syrian glass. The second was Eugène Rousseau, a commissioning dealer in ceramics who had turned to glasswork at the end of the 1860s and was at the height of his achievement in the years *c.* 1880. Typically his glasses were thick walled and translucid, often with interior crackling and shot with random streaks of colour. In 1885 he associated with E. Léveillé, who continued to work in a similar style after Rousseau's death in 1891. The third of the individual artists at the 1878 exhibition and the best known of them was Émile Gallé of Nancy, who had been experimenting in glasswork since about 1867. His earliest work was in clear glass, lightly tinted and decorated with enamel and engraving. But he soon developed the use of deeply coloured, almost opaque glasses in heavy masses, often layered in several thicknesses and carved or etched to form plant motifs. His work reflected the prevailing interest in Japanese art and with its frequently asymmetrical form contributed largely to the Art Nouveau of the end of the century. In this period much of Gallé's manner was reflected in the glassware produced on a more commercial basis by the firm of Daum Frères of Nancy.

Gallé glass

A number of French artists successfully explored the use of *pâte de verre* (powdered glass fired in a mold). The pioneer in its use was Henri Cros, who was working near the end of the 19th century. It was later the medium for important work by Albert Dammouse and François Décorchemont.

Among the later leaders of French glass art was René Lalique, who around the 1920s was producing his most typical work, which is characterized by relief decoration produced by blowing into molds or by pressing. He was a leading advocate of the use of glass in architecture and much of his work was in the form of lighting equipment and in details of interior decoration. The work of his contemporary, Maurice Marinot, was more in the tradition of Rousseau, with heavy, thick-walled vessels in strong forms often with boldly cut-away abstract decoration; and Henri Navarre in the 1930s was producing work of a similar monumental nature.

The most significant work of Jean Luce and Marcel Goupy, designers of glass and ceramics, was in the production of elegant tablewares. For a long period André Thuret made glasses in thick plastic forms; and Jean Sala worked in bubbled glass. The firm of Daum was distinguished, after World War II, by its thick clear glass vessels manipulated into flowing shapes to designs by Michel Daum.

THE SCANDINAVIAN COUNTRIES

Development of modern Swedish glass

Up to the time of World War I the Swedish glass industry produced little original work. The sudden development of modern Swedish glass in the 1920s was attributable mainly to the initiative of the Swedish Arts and Crafts Society that resulted in the employment of the painters Simon Gate and Edward Hald by Orrefors glassworks and Edvin Ollers by Kosta glassworks, both in the glass-producing area of Småland in southern Sweden. The first results were exhibited in Stockholm in 1917 and consisted of handblown, undecorated tablewares, together with the luxury "Graal" glass with internal stained decoration, which had been rapidly developed under Gate's inspiration at Orrefors. It was, however, engraved glasswork, chiefly that designed by Gate and Hald at Orrefors, on which the reputation of Swedish glass was established in the 1920s and particularly at the Paris International Exhibition of Decorative Arts in 1925.

In the 1930s came a change of direction. The Swedish factories began to take less interest in engraving and followed the initiative of the French artists in making thick tinted and figured glasses. In this mode they found their greatest success—attributed largely to their having achieved a system of intimate association between the artists and the glassmaker craftsmen.

At Orrefors additional artists were added to the establishment from 1929 onward, including Vicke Lindstrand, Sven Palmqvist, Nils Landberg, Edvin Öhrström, John Selbing, and Ingeborg Lundin. Each of them worked in an individual style, and in addition to decorative pieces many of them designed tablewares for the subsidiary Sandvik factory. At Kosta important work was produced by Elis Bergh and later by Lindstrand. Gerda Strömberg designed for both Eda glassworks and for Strömbergshyttan. In the 1960s many new methods of forming and decorating glass were explored by young designers; and an element of the current Pop art was discernible, such as in the work of Gunnar Cyrén at Orrefors.

In Denmark the Holmegaard glassworks and in Norway the Hadeland glassworks both followed in some respects the example of Swedish glass. At Holmegaard the movement began in the late 1920s with the appointment as art director of Jacob E. Bang, whose designs included an amount of striking engraved work, and was continued in the clean forms of his successor, Per Lütken. At Hadeland some distinctive glass was designed by a number of artists including Sverre Pettersen, Willy Johansson, and Arne Jon Jutrem.

By courtesy of Die Neue Sammlung, Munich



Figure 231: Double-cased glass vases designed by Timo Sarpaneva, Iittala glassworks, Finland, 1957. In Die Neue Sammlung, Munich. Height (left) 30 cm., (right) 17.5 cm.

Modern Finnish design

In Finland original modern work of great significance has been carried out. Following the example of the Swedish factories, the artist Henry Ericsson was appointed designer at the Riihimäki glassworks in the late 1920s, and Göran Hongell was employed in a similar capacity at the Karhula glassworks in the 1930s. At this time the well-known Finnish artists Arttu Brummer and Alvar Aalto were also concerned in glass design. Shortly after World War I the influential designer Gunnel Nyman was producing glasses freely blown in thick masses to form asymmetrical shapes. Other important designers were Tapio Wirkkala and Timo Sarpaneva working for the Iittala glassworks (Figure 231), Kaj Franck for the Nuutajärvi glassworks (trading as Wärtsilä-Notsjö), and Helena Tynell and Nanny Still for Riihimäki. In the 1960s Timo Sarpaneva struck a new note with his sculptures formed from the charred inner surface of wooden molds, while Oiva Toikka designed for Wärtsilä-Notsjö objects of a markedly Pop art nature.

BELGIUM AND THE NETHERLANDS

In Belgium the Val-Saint-Lambert factory was an important producer of heavily cut crystal throughout the period. It is also associated with layered work and was particularly prominent with original work of this nature around 1900. Later Charles Graffart designed for it wares made in a variety of techniques, some of them with engraved decoration.

The Dutch glassworks at Leerdam played an important part in the modern movement and followed a line of development distinct from that of the Scandinavian factories. In 1915 the decision was made to invite designs from artists, and by the early 1920s excellent simple tablewares were being made to designs by the architects K.P.C. de Bazel and H.P. Berlage and by the decorative artist C. de Lorm. From the early 1920s onward individually designed pieces called Unica were made; some of the earlier examples were by Chris Lebeau, but most were produced by Andries D. Copier. Later decorative work included designs by Floris Meydam and Willem Heesen.

ITALY

Variations of traditional Venetian glass

By the middle of the 19th century, Italian glassmaking had partly revived. In the 1860s the Museo Vetrario was founded at Murano (Venice), and Antonio Salviati began to produce the glasses that attracted much attention at the Paris Exhibition of 1867. These were variations of the traditional Venetian style with elaborate furnace decoration, and the production of glasses of this nature continued at Murano throughout the remainder of the 19th century and beyond.

The 1920s saw the development of a more conscious spirit of artistry in Italian glasswork. Paolo Venini was concerned in producing simple elegant glasses designed by the decorative artist Vittorio Zecchin; and G. Balsamo Stella and his Swedish wife Anna were producing engraved work. In later years, both before and after World War II, much research was done in new methods of colouring and figuring glass; the results were seen in the glasses designed by Ercole Barovier for the firm of Barovier & Toso and in those designed by Giulio Radi for the firm Arte Vetraria Muranese.

From the Venini firm, presided over by Paolo Venini until his death in 1959, came many interesting innovations, such as the colourful glasses designed by Carlo Scarpa and by Fulvio Bianconi and an interesting series by the Finn Tapio Wirkkala. For the firm of Vistosi some striking modern glasses were designed by artists such as Peter Pelzel and Alessandro Pianon. Some of the work, such as a series of vases designed by Flavio Poli for Seguso Vetri d'Arte, showed some influence from the thick-glass techniques of the north, but the modern Italian glass mostly retained a distinctly Venetian, volatile character. An experiment of interest was the production of a series of glass

sculptures from sketches and models commissioned by the dealer Egidio Constantini from internationally prominent painters and other artists. (Hu.Wa.)

Chinese glass

Glass has never been truly at home in China. Records suggest that it was brought there from the West as early as the 3rd century, but finds of small glass objects of typical Chinese shapes dating from as early as the Han dynasty (206 BC–AD 220) suggest that, even if the material was brought from the West, it could be worked on the spot to conform to Chinese usage. It was no doubt regarded as a cheap substitute for jade. The Chinese themselves do not claim to have made glass before the 5th century, and even then it is doubtful if they knew more than how to make beads and other similar small objects. The vessels of glass occasionally found in burials of the T'ang (618–907) and later dynasties, although perhaps locally made, are more likely imports. Of the extant glass vessels typically Chinese in form, none can be shown to be of a date earlier than the reign of the K'ang-hsi emperor (1661–1722), and there is every likelihood that glassmaking was in fact introduced in this period when, through the Jesuits, China became vividly aware of Western culture. To this period probably belongs a series of bowls and vases of which the blown character is manifest. They are often of a deteriorated material that appears to suffer from the same defects as European glass of the same epoch.

During the reigns of the Yung-cheng (1722–35) and Ch'ien-lung (1735–96) emperors, the emphasis on blown forms is subordinated to the desire to make glass a surrogate for natural stones. Although the colours used are often not such as are found in nature, the glass is handled as though it were jade, the foot in particular being fash-

By courtesy of Museum für Kunst und Gewerbe, Hamburg



Figure 232: Snuff bottle, opaque whitish glass with red cut overlay, Chinese, 18th century. In the Museum für Kunst und Gewerbe, Hamburg. Height 10.5 cm.

ioned as though cut from stone. This lapidary treatment is further emphasized in the cased glass bottles cut on the wheel in such a way that the design stands in one or more colours on a ground of a contrasting tone (Figure 232). (R.J.Ch.)

BIBLIOGRAPHY

Interior design. *General works:* ARNOLD FRIEDMANN, JOHN F. PILE, and FORREST WILSON, *Interior Design: An Introduction to Architectural Interiors* (1970), an introduction to the field of interior architecture written for students of design; SHERRILL WHITON, *Elements of Interior Design and Decoration*, 3rd ed. (1963), a scholarly text; RAY and SARAH FAULKNER, *Inside Today's Home*, 3rd ed. (1968), a thorough and well-illustrated book on the interior design of homes; DIANA ROWNTREE, *Interior Design* (1964), a brief and personal view of interior design written primarily for British readers; EDGAR KAUFMAN, *What Is Modern Interior Design?* (1953, reprinted 1969), a very brief but perceptive treatise. A later monograph on home decorating is MARY GILLIAT, *The Decorating Book* (1981), with special photography by Michael Dunne.

Special types of interiors: MICHAEL SAPHIER, *Office Planning and Design* (1968), a clear overview of the field of business and office interiors; BETTY ALSWANG and AMBUR HIKEN, *The Personal House* (1961), a photographic collection of very personal interiors primarily designed by the artist-occupants, rather than professional interior designers. The photographs and comments contained in the following works make them significant sources for the study and understanding of special interiors: WILLIAM WILSON ATKIN and JOAN ADLER, *Interiors Book of Restaurants* (1960); HENRY END, *Interiors Book of Hotels and Motor Hotels* (1963); JOHN F. PILE, *Interiors Second Book of Offices* (1969); MORRIS KETCHUM, *Shops and Stores*, rev. ed. (1957); GEORGE NELSON (ed.), *Living Spaces* (1952); MARY GILLIATT AND MICHAEL BOYS, *English Style in Interior Decoration* (1967).

Special subjects: JOHANNES ITTEN, *Kunst der Farbe* (1961; Eng. trans., *The Art of Color*, 1961); FABER BIRREN, *Color for Interiors, Historical and Modern* (1963); LESLIE LARSON, *Lighting and Its Design* (1964); JOHN F. PILE (ed.), *Drawings of Architectural Interiors* (1967); MARIO G. SALVADORI and ROBERT HELLER, *Structure in Architecture* (1963), a very readable introduction to structural principles understandable to laymen, but written on a very professional level; MARIO DAL FABBRO, *Modern Furniture*, 2nd ed. (1958), EDWARD LUCIE-SMITH, *The Story of Craft: The Craftsman's Role in Society* (1981), explores the unifying and the distinctive features of craft and fine arts.

Historical developments: GEORGE SAVAGE, *A Concise History of Interior Decoration* (1966), is the only English-language work that summarizes the history of the subject. Information about the earliest furniture may be found in HOLLIS S. BAKER, *Furniture in the Ancient World* (1966); and the *Natural History* (various editions) of PLINY THE ELDER, which contains much information in the final volumes on the Roman scene. Books about the Middle Ages are not numerous, but the *Guide to the Early Christian and Byzantine Antiquities* of the British Museum (1921), is a useful work. There are many books dealing with various aspects of the Renaissance, such as PETER and LINDA MURRAY, *The Art of the Renaissance* (1963); FRIDA SCHOTTMULLER, *Furniture and Interior Decoration of the Italian Renaissance* (1921); PIERRE DU COLOMBIER, *Le Style Henri IV–Louis XIII* (1941); GERMAIN BAZIN, *Classique, baroque et rococo* (1964; Eng. trans., *Baroque and Rococo*, 1964); and VICTOR TAPIE, *Baroque et classicisme* (1957; Eng. trans., *The Age of Grandeur*, 1960). PIERRE VERLET, *Le Mobilier royal français* (1945; Eng. trans., *French Royal Furniture*, 1963) and *Les Meubles français du XVIII siecle* (1956; Eng. trans., *French Furniture and Interior Decoration*, 1967), are important works by a great authority dealing with 18th-century developments. GEORGE SAVAGE, *French Decorative Art, 1638–1793* (1969), discusses most of the objects in general use for interior decoration. FISKE KIMBALL, *The Creation of the Rococo* (1943), is an important examination of the sources of this style; TERISIO PIGNATTI, *Il Rococo* (1967, Eng. trans., *The Age of Rococo* 1967; Eng. trans., *The Age of Rococo*, 1969), is a scholarly picture-book based on an exhibition so titled. ADRIEN FAUCHIER-MAGNAN, *Les Petites Cours d'Allemagne au XVIII Siecle* (1947; Eng. trans., *Small German Courts in the Eighteenth Century*, 1958), is valuable for information about the pervasion of French art and culture. The Wallace Collection (London) catalog of *Furniture* by F.J.B. WATSON (1956), and the catalog of *Sculpture* by JAMES G. MANN (1931), are scholarly works essential to the study of their subject; see also F.J.B. WATSON, *Louis XVI Furniture* (1960).

On English decoration the works of MARGARET JOURDAIN: *English Decoration and Furniture of the Early Renaissance* (1924), *Regency Furniture, 1795–1820* (1934), and *The Work of William Kent* (1948), are all worth consulting. PERCY MACQUOID and RALPH EDWARDS, *The Dictionary of English Furniture from the Middle Ages to the Georgian Period*, 2nd ed. rev., 3 vol. (1954), is a scholarly and important work. THOMAS A. STRANGE, *English Furniture, Decoration, Woodwork, and Allied Arts* (1900; reprinted 1950), reproduces many pages from 18th-century English design books, including Chippendale's *Director*; and HUGH HONOUR, *Neo-Classicism* (1968), discusses the style in its international implications, as well as dealing with that of the brothers Adam. SUSAN LASDUN, *Victorians at Home* (1981), discusses domestic interior in the period from 1820 to 1900.

For the Gothic revival there is no better source than SIR KENNETH CLARK, *The Gothic Revival* (1928, reprinted 1970). J. MORDAUNT CROOK, *William Burges and the High Victorian Dream* (1981), is a study of the life of a protagonist of the Gothic revival style in design. JOSEPH DOWNS, *American Furniture* (1952), is a standard work. GEORGE SAVAGE, *The Dictionary of Antiques* (1970), discusses former objects of interior decoration and their style from the Renaissance onwards. Art Nouveau has been the subject of a number of books in recent years. Among the best are MARIO AMAYA, *Art nouveau* (1966); MARTON BATTERSEY, *The World of Art Nouveau* (1968); and STEPHEN TSCHODI MADSEN, *Source of Art Nouveau* (1956).

There are no works discussing Oriental interior decoration only, and information must, for the most part, be gleaned from books discussing specific types of objects, such as painting, porcelain, furniture, and bronze. By far the best source is *Chinese Art*, 4 vol. (1960–65), an international symposium by several well-known Orientalists that discusses almost everything of importance to the subject. A useful general survey is LEIGH ASHTON (ed.), *Chinese Art*, by several well-known authorities (1935), which summarizes in one volume the salient facts about works in many differing materials. For Japanese art, MARCUS B. HUISH, *Japan and Its Art* (1889), is an excellent general work, but few books that can be recommended for the present purpose have been published in English.

On Islāmic art, an excellent work is DAVID TALBOT RICE, *Islamic Art* (1965); a more detailed survey is MAURICE S. DIMAND, *A Handbook of Muhammadan Art*, 3rd ed. rev. (1958). A.U. POPE and PHYLLIS ACKERMAN, *A Survey of Persian Art from Prehistoric Times to the Present*, 14 vol. (1938–67), should be consulted for this aspect. FRANZ BOAS, *Primitive Art*, new ed. (1955), discusses the principles behind the decoration of a wide variety of art of this kind, with special attention to the North Pacific Coast of North America. ERIC LARRABEE and MASSIMO VIGNELLI, *Knoll Design* (1981), is a history of modern commercial interior design.

Furniture and accessory furnishings. OLE HANSCHER, *Møbelkunsten* (1966; Eng. trans., *The Art of Furniture*, 1967), chronological survey of the art of furniture (with illustrations). GEORGE NAKASHIMA, *The Soul of a Tree: A Woodworker's Reflections* (1981), is an inspirational commentary of a designer and crafter on the art of furniture making.

Near East and classical antiquity: A. LUCAS, *Ancient Egyptian Materials and Industries*, 3rd ed. (1948), indispensable; HOWARD CARTER and A.C. MACE, *The Tomb of Thut-Ankh-Amen*, 3 vol. (1923–27); GISELA H. RICHTER, *The Furniture of the Greeks, Etruscans and Romans* (1966), the standard reference work in this area.

Middle Ages: VIOLLET-LE-DUC, *Dictionnaire raisonné bilier français de l'epoque carlovingienne a la Renaissance*, 6 vol. (1858–75), still an authoritative work.

Renaissance and later: (*Italy*): GEORGE LELAND HUNTER *Italian Furniture and Interiors*, 2 vol. (1918), mostly illustrations; WILLIAM M. ODOM, *A History of Italian Furniture from the 4th to the Early 19th Centuries*, 2 vol. (1918–19). (*Spain*): ARTHUR BYNE and MILDRED STAPLEY, *Spanish Interiors and Furniture* (1921), profusely illustrated with scale drawings and photographs. (*Germany*): HEINRICH KREISEL, *Die Kunst des deutschen Möbels*, 2 vol. (1968–70), thorough, illustrated history of German furniture. (*France*): PIERRE VERLET, *Le Mobilier royal français*, 2 vol. (1945–55); *Les Meubles français du XVI-IIe siecle*, 2 vol. (1956), a learned treatise on French furniture. (*England and the colonies*): PERCY MACQUOID and RALPH EDWARDS, *The Dictionary of English Furniture from the Middle Ages to the Late Georgian Period*, 2nd ed., 3 vol. (1954), documented survey of English and American furniture; RALPH FASTENEDGE, *English Furniture Styles from 1500 to 1830* (1962), an excellent elementary introduction to the study of English furniture; ANTHONY COLERIDGE, *Chippendale Furniture* (1968), illustrated study of Chippendale and his contemporaries; CLIFFORD MUSGRAVE, *Adam and Hepplewhite and Other Neo-Classical Furniture* (1966), written by one of the best informed students of the Neoclassical English style of furniture; CHARLES F. MONTGOMERY, *American Furniture* (1966), a survey of Federal period furniture. See also BERRY B. TRACY, *The Federal Furniture and Decorative Arts at Boscobel* (1981); CHARLES SANTORE, *The Windsor Style in America* (1981); JOHN T. KIRK, *American Furniture and the British Tradition to 1830* (1983).

19th century and modern: E.D. and F. ANDREWS, *Shaker Furniture: The Craftsmanship of an American Communal Sect* (1937); R.V. SYMONDS and B.B. WHINERAY, *Victorian Furniture* (1962), with many illustrations; SERGE GRANDJEAN, *Empire Furniture, 1800 to 1825* (1966); MIKOLAUS PEVSNER, *Pioneers of Modern Design from William Morris to Walter Gropius* (1960), JEAN CASSOU, EMILE LANGUE, and NIKOLAUS PEVSNER, *Les Sources du vingtieme siecle* (1961); JOHN F. PILE, *Modern Furniture* (1979); JONATHAN L. FAIRBANKS and ELIZABETH BIDWELL BATES, *American Furniture, 1620 to the Present* (1981); DAVID A. HANKS, *Innovative Furniture in America from 1800 to the Present* (1981).

Rugs and carpets. For further study in this area the following summaries are recommended: FRIEDRICH SARRE and HERMANN TRENKWALD, *Altorientalische Teppiche*, 2 vol. (1926–28; Eng. trans., *Old Oriental Carpets*, 2 vol., 1926–29), excellent large plates with exact description and technical analysis of masterpieces of Oriental rugs and carpets, including a comprehensive bibliography; KURT ERDMANN, *Der orientalische Knüpfteppich*, 3rd ed. (1965; Eng. trans., *Oriental Carpets*, 2nd ed., 1962), an important discussion of Oriental rugs and carpets with regard to artistic development—contains a comprehensive bibliography for all areas of the Oriental carpet, including newspaper and journal essays, and museum and auction catalogs, arranged according to subject areas; and *Siebenhunderi Jahre Orientteppich* (1966; Eng. trans., *Seven Hundred Years of Oriental Carpets*, (1970), a posthumous collection of the author's articles, presenting material not found in his earlier works; WILHELM VON BODE and ERNST KUHNEL, *Vorderasiatische Knüpfteppiche aus alter Zeit*, 4th ed. (1955; Eng. trans., *Antique Rugs from the Near East*, 4th rev. ed. 1970), a handbook on the scientific study of rugs and carpets; ALBERT F. KENDRICK and C.E.C. TATTERSALL, *Handwoven Carpets, Oriental and European*, 2 vol. (1922, reprinted 1973), a monographic treatment of the individual types, including products after 1800; ARTHUR V. POPE and PHYLLIS ACKERMAN, *The Art of Carpet Making in a Survey of Persian Art from Prehistoric Times to the Present*, 7 vol. (1938–39), the most comprehensive study of the art of carpet making in Persia, richly illustrated; CORNELIA B. FARADAY, *European and American Carpets and Rugs* (1929), a comprehensive study of European and American carpet production, including native arts and machine-made carpets (very richly illustrated but without bibliography); C.E.C. TATTERSALL, *A History of British Carpets*, rev. ed. (1966), an extensive study of carpet production in England, including machine-made carpets; M.J. MAJORCAS, *English Needlework Carpets, 16th–19th Centuries* (1963), a richly illustrated treatise; MADELEINE JARRY, *Manufacture nationale de la Savonnerie* (Eng. trans., *The Carpets of the Manufacture de la Savonnerie* 1966) and *The Carpets of Aubusson* (1969), two studies of French floor coverings, with ample illustrations and VALERIE JUSTIN *Flat-woven Rugs of the World: Kilim, Soumak, and Brocading* (1980). See also P.R.J. FORD, *The Oriental Carpet: A History and Guide to Traditional Motifs, Patterns, and Symbols* (1981); and GIOVANNI CURATOLA, *The Simon and Schuster Book of Oriental Carpets* (1982).

Tapestry. W.G. THOMSON, *A History of Tapestry from the Earliest Times to the Present Day*, 3rd ed. rev. and ed. by F.P. THOMSON and E.S. THOMSON (1973), a standard work on the history of tapestry, which has been updated by F.P. THOMSON, *Tapestry: Mirror of History* (1980); M.J. GUIFFREY, E. MUNTZ, and A. PINCHART, *Histoire generale de la tapisserie*, 3 vol. (1878–85), French tapestries discussed by Guiffrey, Italian tapestries by Muntz, Flemish tapestries by Pinchart; M. FENAILLE, *État general des tapisseries de la manufacture des Gobelins depuis son origine jusqu'a nos jours, 1600–1900*, 6 vol. (1903–23), a work of primary importance, presenting a detailed history of the Gobelins factory; J. BADIN, *La Manufacture de tapisseries de Beauvais, depuis ses origines jusqu'a nos jours* (1909), a basic reference for the history of tapestry production at the Beauvais factory; *Wandteppiche*, 6 vol. (1923–34; Eng. trans. of pt. 1, *Tapestries of the Lowlands*, 1924), a general worldwide treatment of tapestry, with numerous black and white illustrations, although many of the European medieval attributions have been questioned or rejected; G.L. HUNTER, *The Practical Book of Tapestries* (1925), precise and useful descriptions, with numerous reproductions; C.G. JANNEAU, *Evolution de la tapisserie* (1947), illustrations and technical information on collections of European tapestries, some of which have been subsequently disbanded with works relocated; D. HEINZ, *Europäische Wandteppiche*, vol. 1, *Von den Anfängen der Bildwirkerei bis zum Ende des 16. Jahrhunderts* (1963), a thorough treatment of tapestry up to the end of the 16th century, with a typological index, an extensive bibliography, and numerous illustrations; R.A. WEIGERT, *La Tapisserie et le tapis en France* (1964), a scholarly discussion of the history of French tapestry; P. VERLET *et al.*, *La Tapisserie: histoire et technique du 14e au 20e siecle* (1977; Eng. trans., *The Book of Tapestry: History and Technique*, 1978), a well-illustrated volume on Western tapestry from the Middle Ages to the 20th century; MADELEINE JARRY, *La Tapisserie des origines a nos jours* (1968; Eng. trans., *World Tapestry*, 1969), a well-documented study of tapestry throughout the world, including an extensive bibliography and many black and white and colour illustrations, and *La Tapisserie: art du 20e siecle* (1974), a study of the worldwide renascence of tapestry during the 20th century; V. FOUGRE, *Tapisseries de notre temps* (1969), a brief study of contemporary French tapestry, with an index of tapestry artists and illustrated with black and white and colour reproductions; R.A. D'HULST, *Flemish Tapestries* (1967), an elaborate study of Flemish tapestry from the Middle Ages to the Baroque periods, with many colour illustrations; R. KAUFMANN, *The New American Tapestry* (1968), a well-illustrated text dealing with technique as well as with the works of leading American tapestry designers and weavers; G. SUTHERLAND, *Coventry Tapestry* (1964), an interesting account of the design, weaving, and installation of Sutherland's tapestry for Coventry cathedral; M.B. FREEMAN, *The Unicorn Tapestries* (1976), a detailed and well-illustrated history of these tapestries, which are housed at the Cloisters, a branch of the Metropolitan

Museum of Art, New York City; P. ACKERMAN, *Tapestry: The Mirror of Civilization* (1933, reprinted 1970), one of the classic works in English dealing with the historical development of European tapestry. LAYA BROSTOFF, *Weaving a Tapestry* (1982), is a brief overview with bibliography.

Collection catalogs: Important illustrated catalogs of tapestry collections include: A.S. CAVALLO, *Tapestries of Europe and Colonial Peru in the Museum of Fine Arts, Boston*, 2 vol. (1967); D. DU BON, *Tapestries from the Samuel H. Kress Collection at the Philadelphia Museum of Art* (1964); H.C. MARILLIER, *The Tapestries at Hampton Court Palace, London*, 2nd ed. (1962), and with A.J. WACE, *Marlborough Tapestries at Blenheim Palace* (1968); N.Y. BIRYUKOVA, *The Leningrad Hermitage Gothic and Renaissance Tapestries* (1966); M. CRICK KUNTZIGER, *Musees royaux d'Art et d'Histoire: Catalogues des tapisseries* (1956); A.M. ERKELENS, *Wandtapijten I. Late Gotetk en vroege Renaissance, Rijksmuseum, Amsterdam* (1962); and A.F. KENDRICK, *Catalogue of Tapestries at the Victoria and Albert Museum* (1924). Tapestries of the second half of the 20th century are shown in numerous exhibition catalogs, such as ERIKA SELLMAN-BÜSCHING, *Tapisserien 1970 bis 1981* (1981); F. VIALET, *Tapisseries contemporaines d'Aubusson* (1981); *Swiss Tapestries, Artists of Today* (1981).

Floral decorations. LIBERTY HYDE BAILEY and ETHEL ZOE BAILEY (comps.), *Hortus Second: A Concise Dictionary of Gardening, General Horticulture and Cultivated Plants in North America* (1941), basic for nomenclature; VICTOR LORET, *La Flore pharaonique d'apres les documents hieroglyphiques et les specimens decouverts dans les tombes*, 2nd ed. (1892), includes information concerning wreaths and garlands; CHARLES VICTOR DAREMBERG and E. SAGLIO, *Dictionnaire des antiquites Grecques et Romaines d'apres les textes et les monuments*, vol. 1, pt. 2 (1877), lists flowers grown and ornamental uses (under "Corona" and "Coronarius et Coronaria"); JOHN GERARD, *The Herball* (1597), descriptions and contemporary wood engravings of English garden flowers; JOHN PARKINSON, *Paradisi in Sole Paradisus Terrestris* (1629), descriptions and usage of flowers in 17th-century England; P. GIOVANNI BATTISTA FERRARI, *Flora ouero cultura di fiori* (1633), on the culture and care of cut flowers, including how to preserve, arrange, and ship them, with interesting illustrations; PHILIP MILLER, *The Gardeners Dictionary*, 2 vol. (1735), an important and popular 18th-century work, with full descriptions of garden flowers and illustrations; HELEN GERE CRUICKSHANN (ed.), *John and William Bartram's America* (1957), contains information about new plant discoveries and exchanges of garden material between America and England in the 18th century; *Godey's Lady's Book* (1830–98), almost monthly advice in the editorial pages about gardening or arranging flowers; J. RAMSBOTTOM, *A Book of Roses* (1939), information about old-fashioned roses; RALPH G. WARNER, *Dutch and Flemish Flower and Fruit Painters of the 17th and 18th Centuries* (1928), profusely illustrated; JULIA S. BERRALL, *A History of Flower Arrangement*, rev. ed. (1968), on all styles and periods, including original source lists of plant materials and many illustrations; MARGARET FAIRBANKS MARCUS, *Period Flower Arrangement* (1952), emphasis on art; JOSIAH CONDER, *The Theory of Japanese Flower Arrangements* (1935), reprint of an original paper read by the author in 1889 to the Asiatic Society of Japan, to which have been added 36 colour plates of Ikenobô and *moribana* arrangements; ALFRED KOEHN, *The Art of Japanese Flower Arrangement (Ikebana): A Handbook for Beginners* (1934), with actual photographs instead of paintings; DONALD RICHIE and MEREDITH WEATHERBY (eds.), *The Masters' Book of Japanese Flower Arrangement: With Lessons by the Masters of Japan's Three Foremost Schools: Sen'ei Ikenobo, Houn Ohara, Sofu Teshigahara* (1966), contains an excellent historic section illustrated from the arts and photographs in colour and black and white contemporary expressions; SHOZO SOTO, *The Art of Arranging Flowers* (1966), on all aspects of Japanese flower arranging, with excellent colour and black-and-white illustrations. Later works include GERTRUDE JEKYLL, *Flower Decoration in the House* (1982), and *Colour Schemes for the Flower Garden*, 8th ed. (1982); EMMA WOOD and JANE MERER, *Flower Crafts* (1982); MARIAN AARONSON, *Flowers in the Modern Manner* (1981); TOKUJI FURUTA, *Interior Landscaping* (1983); INTERIOR PLANTSCAPE ASSOCIATION (U.S.), *Manual of Practice* (1980); MARY ADAMS, *Natural Flower Arranging* (1981); EDITH BLACK, *Modern Flower Arranging* (1982).

Pottery. Books on pottery are fairly numerous, but of those that have been written in recent years for the popular market many are not always reliable and much information is duplicated among them. The following titles may be regarded as standard or major works.

General works: Two good introductory surveys are GEORGE SAVAGE, *Pottery Through the Ages* (1963) and *Porcelain Through the Ages*, 2nd ed. (1963). ROBERT J. CHARLESTON (ed.), *World Ceramics* (1968), is a lavishly illustrated history of ceramics written by many noted specialists in Europe and the United States. Highly recommended is W.B. HONEY, *European Ceramic Art, from the End of the Middle Ages to about 1815*, 2nd ed., 2 vol. (1963), which has a comprehensive list of marks and an excellent bibliography. Also useful is EMIL HANNOVER, *Keramisk Haandbag*, 2 vol. (1919–24; Eng. trans., *Pottery and Porcelain*, 3 vol., 1925); and WARREN COX, *The Book of Pottery and Porcelain* (1944). GEORGE SAVAGE, *The Dictionary of Antiques* (1970), includes information about continental wares on a less comprehensive scale, but its extensive bibliography has been brought up to date. It also differentiates between works in print and those available only in libraries or secondhand. For the technical side of pottery, see BERNARD LEACH, *A Potter's Book* (1940); W.B. HONEY, *The Art of the Potter* (1946); and PAUL RADO, *An Introduction to the Technology of Pottery* (1969). Factory marks are discussed and listed in W.B. HONEY and JOHN P. CUSHION, *Handbook of Pottery and Porcelain Marks*, 3rd rev. ed. (1965); Cushion has also published separate handbooks on *English Ceramic Marks and Those of Wales, Scotland, and Ireland* (1959), on *German Ceramic Marks and Those of Other Central European Countries* (1961), and on *French and Italian Ceramic Marks* (1965). GEOFFREY GOODEN, *Encyclopaedia of British Pottery and Procelain Marks* (1964), is another recommended source. WILLIAM CHAFFERS, *Marks and Monograms on Pottery and Porcelain*, 14th ed. (1931), formerly the standard work, has now been superseded and is sometimes inaccurate. More specialized is ARTHUR BEHSE, *Deutsche Fayencemarken-Brevier* (1955), which should be consulted for German wares. Ceramic terms are defined by GEORGE SAVAGE and H. NEWMAN in the *Illustrated Dictionary of Ceramic Terms* (1973). See also ROBERT FOURNIER, *Illustrated Dictionary of Pottery Form* (1981).

Ancient Near East and Egypt: Books dealing specifically with the pottery of this area and period are scarce. See WALTER ANDRAE (ed.), *Coloured Ceramics from Assur* (1925); and HENRY HALLER, *Egyptian Ceramic Art* (1898). More information will be found in books dealing with the arts of the ancient Near East and Egypt.

Ancient Aegean, Greece, and Italy: Works on ancient pottery from these regions are fairly numerous. A comprehensive selection is listed below. HENRY B. WALTERS, *A History of Ancient Pottery*, 2 vol. (1905, reprinted 1971); CHARLES F. SELTMAN, *Attic Vase-Painting* (1933); VINCENT D'ARBA DESBOROUGH, *Protogeometric Pottery* (1952); G.M.A. RECHFER, *A Handbook of Greek Art*, 6th ed. (1969); *Ancient Italy* (1955); *Attic Red-Figured Vases: A Survey*, 2nd ed. (1958); and *The Craft of Athenian Pottery* (1923); ARTHUR LANE, *Greek Pottery*, 3rd ed. (1971); ROBERT M. COOK, *Greek Painted Pottery* (1960); T.B.L. WEBSTER, *Greek Terracottas* (1951); ROBERT J. CHARLESTON, *Roman Pottery* (1955); JOHN D. BEAZLEY, *Etruscan Vase Painting* (1947). Historical comparisons with modern methods and procedures are the basis of D.P.S. PEACOCK, *Pottery in the Roman World* (1982).

Islámic pottery: There are few works easily available. The best are ARTHUR LANE, *Early Islamic Pottery*, rev. ed. (1958) and *Later Islamic Pottery*, 2nd ed. (1972); see also MAURECE S. DIMAND, *A Handbook of Muhammadan Art*, 3rd ed. rev. (1958); and ROBERT L. HOBSON, *A Guide to Islamic Pottery of the Near East* (1932).

Western pottery: Where they exist, the titles of English works on continental pottery are given below. Those listed in other languages are illustrated works that will supplement the deficiency of works in English. In the case of porcelain, English works that discuss the wares of individual factories are given only in special cases since they are extremely numerous and most general works on the subject include a bibliography. Contemporary European and American developments are well covered in TAMARA PRÉAUD and SERGE GAUTHIER, *Ceramics of the 20th Century* (1982); and RICHARD ZAKIN, *Electric Kiln Ceramics* (1981). (*Spain*) ALICE W. FROTHENGHAM, *Catalogue of Hispano-Moresque Pottery in the Collection of the Hispanic Society of America* (1936); *Talavera Pottery, with a Catalogue of the Collection of the Hispanic Society of America* (1944); *Lustreware of Spain* (1951); ALBERT VAN DER PUT, *Hispano-Moresque Ware of the XVth Century* (1904). (*Italy*) JOSEPH CHOMPRET, *Repertoire de la Mojolique Italienne*, 2 vol. (1949); BERNARD RACKHAM, *Guide to Italian Maiolica* (1933); *Catalogue of Italian Maiolica*, 2 vol. (1940); and *Italian Maiolica* (1952)—Rackham's books are the best source of information in English; ARTHUR LANE, *Italian Porcelain* (1954); GIUSEPPE MORAZZONI, *Le Porcellane Italiane* (1935). (*France and Belgium*): CHARLES DAMIRON, *La Faience artistique de Moustiers* (1919) and *La Faience de Lyon* (1926); JEANNE GIACOMOFFE. *Faiences françaises* (1963; Eng. trans., *French Faience*, 1963); HANS HAUG, *Les Faiences et porcelaines de Strasbourg* (1922); ARTHUR LANE, *French Faience*, 2nd ed. (1970), the best introduction to the subject; FRANCOIS PON-

CEITON and GEORGE SALLES, *Les Poteries Françaises* (1928); PAUL ALFASSA and JACQUES GUERIN *Porcelaine française du XVIIe au milieu du XIXe siecle* (1932); EMILE BOURGEOIS, *Le Biscuit de Sevres au XVIIIe siecle*, 2 vol. (1909); W.B. HONEY, *French Porcelain of the 18th Century* (1950); EUGENE J. SOIL DE MORAINE, *La Manufacture imperiale et royale de porcelain de Tournay*, 3rd ed. of his *Ceramique tournaisienne* by LUCIEN DELPLACE DE FORMANOIR (1937); GEORGE SAVAGE, *Seventeenth and Eighteenth Century French Porcelain* (1960); PIERRE VERLET, SERGE GRANDJEAN, and MARCELLE BRUNET, *Sevres* (1953). *(Germany and Austria)*: KARL KOETSCHAU, *Rheinisches Steinzeug* (1924); HANS MEYER, *Böhmisches Porzellan and Steingut* (1927); GUSTAV E. PAZAUREK, *Deutsche Fayence and Porzellan-Hausmaler*, 2 vol. (1925, reprinted 1970) and *Steingut: Formgebung and Geschichte* (1921); D. RIESEBIETER, *Die deutschen Fayencen des 17. and 18. Jahrhunderts* (1921), the most comprehensive general work on the subject; EDMUND W. BRAUN and JOSEPH FOLHESICS, *Geschichte der K.K. Wiener Porzellan-Manufaktur* (1907); HANS CHRIST, *Ludwigsburger Porzellan-figuren* (1921); JOHN F. HAYWARD, *Viennese Porcelain of the Du Paquier Period* (1952); FRIEDRICH H. HOFMANN, *Frankenthaler Porzellan*, 2 vol. (1911); *Geschichte der Bayerischen Porzellan-Manufaktur Nymphenburg*, 3 vol. (1921–23); and *Das Porzellan-Manufaktur Nymphenburg*, 3 vol. (1921–23); and *Das Porzellan der Europäischen Manufakturen im 18. Jahrhundert* (1932); V.B. HONEY, *Dresden China*, 2nd ed. (1954) and *Germain Porcelain* (1948); GEORG LENZ, *Berliner Porzellan: Die Manufaktur Friedrichs des Grossen, 1763–1786*, 2 vol. (1913); HUGO MORLEY-FLETCHER, *Antique Porcelain in Color: Meissen* (1971); E. POCHE, *Bohemian Porcelain* (n.d.); KARL ROEDER and HICHE OPPENHEIM, *Das Hochster Porzellan* (1925); HAS. SAUERLANDT, *Deutsche Porzellanfiguren des XVIII. Jahrhunderts* (1923); GEORGE SAVAGE, *18th–Century German Porcelain*, 2nd ed. (1967); CHRISTIAN SCHERER, *Das Fürstenberger Porzellan* (1909). *(Switzerland)*: SIEGFRIED DUCRET, *Zürcher Porzellan des 18. Jahrhunderts* (1944). *(The Netherlands)*; CAROLINE H. DE JONGE, *Delft Aardewerk* (1965; Eng. trans., DELFT-CERAMICS, 1969) and *Nederlandse tegels* (1971; Eng. trans., *Dutch Tiles*, 1971); FERRAND W. HUDIG, *Delfter Fayence* (1929); JEAN JUSTICE, *Dictionnaire des marques et monogrammes de la faïence de Delft* (1920; Eng. trans., *Dictionary of Marks and Monograms of Delft Pottery*, (1930); ELISABETH NEURDENBURG, *Old Dutch Pottery and Tiles* (Eng. trans., 1923); BERNARD RACKHAM, *Early Netherlands Maiolica* (1926), a discussion of the earliest pre-Delft wares. *(Scandinavia)*: RICHARD KARSSON, *Die Stralsunder Fayencefabrik, 1757–1790* (1920); ARTHUR HAYDEN, *Royal Copenhagen Porcelain* (1911); ERIK VETTERGREN, *The Modern Decorative Arts of Sweden* (Eng. trans. 1926), (*Russia*): GEORGY LUKOMSKY, *Russisches Porzellan, 1744–1923* (1924); KARVIN, *Russian Porcelains* (1968). *(England)*: HARRY BARNARD, *Chats on Wedgwood Ware* (1924); GEOFFREY BEMROSE, *Nineteenth Century English Pottery and Porcelain* (1952); FREDERICK H. GARNER, *English Delftware* (1948); JOHN E. and EDITH HODGKIN, *Examples of Early English Pottery, Named, Dated and Inscribed* (1891); ELIZA HETEYARD, *The Life of Josiah Wedgwood*, 2 vol. (1865); ERNEST KORTON NANCE, *The Pottery and Porcelain of Swansea and Nantgarn* (1942); W.J. POUNTNEY, *Old Bristol Potteries* (1920); E. STANLEY PRICE, *John Sadler, a Liverpool Pottery Printer* (1949); BERNARD RACKHAM, *Medieval and English Pottery* (1947) and *Early Staffordshire Pottery* (1951); BERNARD RACKHAM and HERBERT READ, *English Pottery: Its Development from Early Times to the End of the Eighteenth Century* (1924); GEORGE V. RHEAD, *The Earthenware Collector* (1920); JOSIAH WEDGWOOD, *Selected Letters*, ed. by ANN FINER and GEORGE SAVAGE (1965); DONALD TOUNER, *Handbook of Leeds Pottery* (1951); FREDERICK WILLIAMSON, *The Derby Pot Manufacturing Known as Cockpit Hill* (1931); HUGH WAKEFIELD, *Victorian Pottery* (1962); PATTERSON D. GORDON PUGH, *Staffordshire Portrait Figures and Allied Subjects of the Victorian Era* (1971); MURIEL ROSE, *Artist-Potters in England* (1955); JOSEPH L. DIXON, *English Porcelain of the Eighteenth Century* (1952); ENGLISH CERAMIC CIRCLE, *English Pottery and Porcelain: Commemorative Catalogue of an Exhibition Held at the Victoria and Albert Museum* (1949); STANLEY W. FISHER, *The Decoration of English Porcelain* (1954) and *English Blue and White Porcelain of the 18th Century* (1947); GEOFFREY GODDEN, *An Illustrated Encyclopaedia of British Pottery and Porcelain* (1966); V.B. HONEY, *Old English Porcelain* (1948); WILLIAM KING, *English Porcelain Figures of the Eighteenth Century* (1925); GEORGE SAVAGE, *18th-Century English Porcelain*, new ed. (1964); CYRIL COOK, *The Life and Work of Robert Hancock* (1948); *William Duesbury's London Account Book, 1751–1753*, with an introduction by MRS. DONALD MCALISTER (1931); WILLIAM H. TAPP, *Jefferyes Hamett O'Neal, 1734–1801* (1938). *(American Indian pottery)*: G.H.S. BUSHWELL and ADRIAN DIGBY, *The Ancient American Pottery* (1955); WOLFGANG HABERLAND, *The Art of North America* (1964). *(United States)*: JOHN RAMSAY, *American Potters and Pottery* (1939); EDWIN A. BARBER, *Tulip Ware of the Pennsylvania-German Potters* (1903); WARREN COK, *The Book of Pottery and Porcelain*, vol. 2 (1944).

Non-Western pottery: (*China*): The best general work is still V.B. HONEY, *The Ceramic Art of China, and Other Countries of the Far East* (1945); but ROBERT L. HOBSON, *Chinese Pottery and Porcelain*, 2 vol. (1915), is often consulted. Hobson's *Wares of the Ming Dynasty* (1923) and *Later Ceramic Wares of China* (1925), are scholarly works, well illustrated for their period. *Chinese Art*, 4 vol. (1960–65), by a number of internationally known experts, covers the whole field of Chinese art, including pottery and porcelain, and is of considerable value to the student. The earliest Chinese wares are discussed in English by CHIN-TING WU in *Prehistoric Pottery in China* (1938). Wares till the Sung dynasty are discussed by BASIL GRAY in *Early Chinese Pottery and Porcelain* (1952); and by ARTHUR L. HETERINGTON in *Early Ceramic Wares of China* (1922) and *Chinese Ceramic Glazes*, 2nd rev. ed. (1948). For wares of the Ming dynasty and subsequently, see ROGER SOAKE JENYNS, *Ming Pottery and Porcelain* (1953) and *Later Chinese Porcelain*, 2nd ed. (1959). See also YUTAKA MINO, *Freedom of Clay and Brush: Through Seven Centuries in Northern China* (1981); MARY TREGEAR, *Song Ceramics* (1982). (*Korea*): G. ST. G.M. GOMPERTZ, *Korean Celadon, and Other Wares of the Koryŏ Period* (1963); and CHEVON. KIM, *The Ceramic Art of Korea* (1961); V. B. HONEY, *Corean Pottery* (1947). (*Japan*): Japanese porcelain is better documented than the pottery of that country. ROGER SOAKE JENYNS, *Japanese Porcelain* (1965), is the definitive work in English on the subject. See also TAOANARE MITSUOKA, *Ceramic Art of Japan* (1949); and ROGER SOAKE JENYNS, *Japanese Pottery* (1971). (*Chinese and Japanese export porcelain*): This includes works on the erroneously termed Oriental Lowestoft, MICHEL BEURDELEY, *Porcelaine de la compagnie des Indes* (1962; Eng. trans., *Porcelain of the East India Companies*; U.S. title, *Chinese Trade Porcelain*, 1962); SIR HARRY M. GARNER, *Oriental Blue and White*, 3rd ed. (1970); discusses early blue-and-white export wares; JOHN A. LLOYD HYDE, *Oriental Lowestoft* (1954); JEAN MCCLURE MUDGE, *Chinese Export Porcelain for American Trade, 1785–1835* (1962); JOHN G. PHILLIPS, *China Trade Porcelain* (1956); WALTER A. STAEHELIN, *The Book of Porcelain* (1966), a book of Chinese illustrations relating to the export trade in the 18th century; SIR A. TUDOR CRAIG, *Armorial Porcelain of the Eighteenth Century* (1925); T. VOLKER, *Porcelain and the Dutch East India Company* (1954), an examination of the records in relation to the import of Chinese and Japanese porcelain in the 17th century.

Basketry. H.H. BOBART, *Basketwork Through the Ages* (1936, reissued 1971); M.L. LEE, *Basketry and Related Art* (1948); G.M. CROWFOOT, "Textiles, Basketry and Mats," in *A History of Technology*, vol. 1 (1954); GEOFREY H.S. BUSHNELL, "Basketry," *Encyclopedia of World Art*, vol. 2, col. 387–400 (1960); H. BALFET, "La Vannerie, essai de classification," *L'Anthropologie* (1952; Eng. trans. and preface by M.A. BAUMHOFF, "Basketry: A Proposed Classification," in *Papers on California Archaeology*, no. 47–49, pp. 1–21, 1957), a detailed explanation the classifications used in this article and the source of portions adapted for use here with permission of the Archaeological Survey, University of California, Berkeley; O.T. MASON, "Aboriginal American Basketry," *Annual Report, 1902* of the U.S. National Museum, pp. 171–548 (1904), a classic work on basketry; H. MUNSTERBERG, *The Folk Arts of Japan*, ch. 3 (1958); ED ROSSBACH, *Baskets as Textile Art* (1973), on the aesthetics of basket design; GLORIA ROTH TELFEKI, *The Baskets of Rural America* (1957), a history of the origins and techniques of basketry. See also SHEREEN LAPLANTZ, *Pleated Basketry: The Woven Form* (1982); JOHN RICE IRWIN, *Baskets and Basket Makers in Southern Appalachia* (1982).

Metalwork. GEORGIUS AGRICOLA, *De re metallica* (1556; Eng. trans., 1912, reprinted 1950), a scholarly translation of a mining and metallurgical classic; LESLIE AITCHISON, *A History of Metals*, 2 vol. (1960), outstanding for its completeness, competence, and excellent index; *The Pirotechnia of Vannoccio Biringuccio*, trans. from the Italian with introduction and notes by CYRIL STANLEY SMITH and MARTHA TEACH GNUDI (1942), a description of Biringuccio's practices of smelting and metallurgy; HERBERT H. COGHLAN, *Notes on the Prehistoric Metallurgy of Copper and Bronze in the Old World* (1951), an authoritative study with a chapter on various methods of working, such as forging, casting, and sheet metalworking; ROBERT J. FORBES, *Metallurgy in Antiquity*, 9 vol. (1950; new ed., *Studies in Ancient Technology*, 1964–), vol. 8 devoted to the discussion of early metallurgy, the smith and his tools, gold, silver and lead, zinc and brass, and vol. 9 containing the chapters on copper, tin and bronze, and iron—these publications are authoritative and the bibliographies are comprehensive; HANNS U. HAEDEKE, *Metalwork* (1970), a study of European metalwork from the Middle Ages to the 19th century that emphasizes the socio-economic aspects of decorative arts in copper, brass, bronze, iron, and pewter; R. GOOWDWIN-SMITH,

English Domestic Metalwork (1937), deals with technology and style as well as types of domestic objects and utensils; RAYMOND LISTER, *The Craftsman in Metal* (1966), an enlightening discussion of the techniques of metalworking in various historical periods; THOMAS A. RICKARD, *Man and Metals: A History of Mining in Relation to the Development of Civilization*, 2 vol. (1932), shows that civilization was developed by the skilful use of metals in industry and the arts; CHARLES SINGER *et al.* (eds.), *A History of Technology*, 5 vol. (1954–58), the standard general reference book in the field of technology that covers the history of metalwork from early times to about 1900 AD—each subject is written by a master, the illustrations are numerous, well selected, and well explained; R.F. TYLECOTE, *Metallurgy in Archaeology: A Prehistory of Metallurgy in the British Isles* (1962), includes chapters on gold, copper and copper alloys, tin and tin alloys, lead and silver, methods of fabrication, and cites extensive references (the last half of the book is devoted to the study of iron). Later studies include JAMES A. MULHOLLAND, *A History of Metals in Colonial America* (1981); and OPPI UNTRACHT, *Jewelry Concept and Technology* (1982).

Silver and gold: (*Western*): STATON ABBEY, *The Goldsmith's and Silversmith's Handbook*, 2nd ed. rev. (1968); P. ACKERMAN, "The Art of the Parthian Silver- and Goldsmiths," E. MARGULIES, "Colsomne Enamel," and J. ORBELI, "Sasanian and Early Islamic Metalwork," in *A Survey of Persian Art*, ed. by A.U. POPE, vol. 1 (1938); LAWRENCE ANDERSON, *The Art of the Silversmith in Mexico, 1519–1936*, 2 vol. (1941); CLARA LOUISE AVERY, *Early American Silver* (1930, reprinted 1968); GUDMUND BOESEN AND CHRISTEN A. BOJE, *Gammelt dansk sølv til bordbrug* (1948; Eng. trans., *Old Danish Silver*, 1949); KATHRYN C. BUHLER, *American Silver* (1950); BENVENUTO CELLINI, *Treatises . . . on Goldsmithing and Sculpture* (Eng. trans. 1898, reprinted 1966); MICHAEL CLAYTON, *The Collector's Dictionary of the Silver and Gold of Great Britain and North America* (1971); ERNEST M. CURRIER, *Marks of Early American Silversmiths . . .* (1938, reprinted 1970); FRANK DAVIS, *French Silver* (1970); ERIC DELIEB, *Investing in Silver*, new ed. (1970); FAITH DENNIS, *Three Centuries of French Domestic Silver*, 2 vol. (1960); JOHAN W. FREDERIKS, *Dutch Silver*, 4 vol. (1952–61), Renaissance–18th century; JOHN F. HAYWARD, *Huguenot Silver in England, 1688–1727* (1959); HENRY D. HILL, *Antique Gold Boxes* (1953); GRAHAM HOOD, *American Silver: A History of Style, 1650–1900* (1971); G.E.P. and J.P. HOW, *English and Scottish Silver Spoons*, 3 vol. (1952); G. BERNARD and THERLE HUGHES, *Three Centuries of English Domestic Silver, 1500–1820* (1968); G. BERNARD HUGHES, *Small Antique Silverware* (1957); CHARLES J. JACKSON, *English Goldsmiths and Their Marks*, 2nd ed. rev. (1921, reprinted 1964); HEINZ LEITERMANN, *Deutsche Goldschmiedekunst* (1953); Y. OKADA, "History of Japanese Ceramics and Metalwork," *Pageant of Japanese Art*, vol. 4 (1952); CHARLES C. OMAN, *English Domestic Silver*, 6th ed. (1965); JOHN MARSHALL PHILLIPS, *American Silver* (1949); JONATHAN STONE, *English Silver of the Eighteenth Century* (1965); GERALD TAYLOR, *Silver*, rev. ed. (1964) and *Continental Gold and Silver* (1967); PATRICIA WARDLE, *Victorian Silver and Silver-Plate* (1963). (*Modern*): ESBJORN HIORT, *Modern Danish Silver* (1954); GEORG JENSEN, INC., *Fifty Years of Danish Silver in the Georg Jensen Tradition* (1954); WORSHIPFUL COMPANY OF GOLDSMITHS, *Modern British Silver* 1951. 1954, 1959, 1964). (*Middle and Far East*): HENRY L. ROTH, *Oriental Silver, Malay and Chinese* (1910, reprinted 1966); HARRY L. TILLY, *The Silverwork of Burma* (1902). (*North and South America—Pre-Columbian*): JOHN ADAIR, *The Navajo and Pueblo Silversmiths* (1944, reprinted 1970); JOSE PEREZ DE BARRADAS, *Orfebrería prehispánica de Colombia*, 4 vol. (1954–58); ALFONSO CASO, "La Orfebrería prehispánica," in *Artes de Mexico*, no. 10 (1955); DUDLEY T. EASBY, JR., "Ancient American Goldsmiths," *Natural History*, 65:401–409 (1956); MARSHALL H. SAVILLE, *The Goldsmith's Art in Ancient Mexico* (1920); ARTHUR S. WOODWARD, *A Brief History of Navajo Silversmithing* (1938). (*Sheffield plate and pewter*): FREDERICK BRADBURY, *British and Irish Silver Assay Office Marks, 1544–1968 . . .*, 12th ed. (1968); HOWARD HERSCHEL COTTERELL, *Pewter Down the Ages*, 2 pt. (1932) and *Old Pewter: Its Makers and Marks in England, Scotland and Ireland* (1929, reprinted 1963); JOHN B. KERFOOT, *American Pewter* (1924); H.J.L.J. MASSE, *Chats on Old Pewter*, ed. and rev. by RONALD F. MICHAELIS (1949); EDWARD WENHAM, *Old Sheffield Plate* (1955); SEYMOUR B. WYLER, *The Book of Sheffield Plate, with All Known Makers' Marks Including Victorian Plate Insignia* (1949).

Ironwork: MAXWELL AYRTON and ARNOLD SILCOCK, *Wrought Iron and Its Decorative Use* (1929); ARTHUR and MILDRED S. BYNE, *Spanish Ironwork* (1915); HERBERT H. COGHLAN, *Notes on Prehistoric and Early Iron in the Old World* (1956); CHARLES J. FFOULKES, *Decorative Ironwork from the XIth to the XVIIIth Century* (1913); EDGAR B. FRANK, *Petite Ferronnerie ancienne* (1948; Eng. trans., *Old French Ironwork*, 1950); J. STARKIE GARDNER, *English Ironwork of the XVIIth and XVIIIth Centuries* (1911) and *Continental Ironwork of the Renaissance and Later Periods*, rev. ed. (1930); GERALD K. GEERLINGS, *Wrought Iron in Architecture* (1929) and *Metal Crafts in Architecture* (1929); JOHN GLOAG and DEREK BRIDGWATER, *A History of Cast Iron in Architecture* (1948); JOHN HARRIS (comp.), *English Decorative Ironwork from Contemporary Source Books, 1610–1836* (1960); OTTO HOVER, *Das Eisenwerk*, 3rd rev. ed. (1953; Eng. trans., *A Handbook of Wrought Iron from the Middle Ages to the End of the Eighteenth Century*, U.S. title, *Wrought Iron: Encyclopedia of Ironwork;* 1962); J. SEYMOUR LINDSAY, *Iron and Brass Implements of the English House*, rev. ed. (U.S. title, *Iron and Brass Implements of the English and American House*; 1964); RAYMOND LISTER, *Decorative Wrought Ironwork in Great Britain* (1957) and *Decorative Cast Ironwork in Great Britain* (1960); JOSEPH NEEDHAM, "Iron and Steel Production in Ancient and Medieval China," in *Clerks and Craftsmen in China and the West*, ch. 8 (1970); WALLACE NUTTING, *Early American Ironwork* (1919); ALBERT H. SONN, *Early American Wrought Iron*, 3 vol. (1928).

Leadwork: WILLIAM R. LETHABY, *Leadwork, Old and Ornamental, and for the Most Part English* (1893); SIR LAWRENCE WEAVER, *English Leadwork: Its Art and History* (1909); GEORGE ZARNECKI, *English Romanesque Lead Sculpture: Lead Fonts of the Twelfth Century* (1957).

Copper, brass, and bronze: FREDERICK BURGESS, *Chats on Old Copper and Brass*, rev. ed. (1954); HENRY J. KAUFFMANN, *American Copper and Brass*, (1968); ALBERT J. KOOL, *Early Chinese Bronzes* (1970); HERMANN LEISINGER, *Romanesque Bronzes* (*op. cit.*); DAVID G. MITTEN and SUZANNAH F. DOERINGER, *Master Bronzes from the Classical World* (1968); HUGO MUNSTERBERG, *Chinese Buddhist Bronzes* (1967); *Macklin's Monumental Brasses*, rev. by JOHN PAGE-PHILLIPS (1969); JOHN T. PERRY, *Dinanderie: A History and Description of Mediaeval Art Work in Copper, Brass, and Bronze* (1910); JOHN POPE-HENNESSEY, *Renaissance Bronzes from the Samuel H. Kress Collection* (1965); GEORGE SAVAGE, *A Concise History of Bronzes* (1968); C. SIVARAMAMURTI, *South Indian Bronzes* (1963); ERNEST R. SUPFLING, *English Church Brasses: From the 13th to the 17th Century* (1970); ALEXANDER SOPER, *Chinese, Korean, and Japanese Bronzes* (1966); LEON UNDERWOOD, *Bronzes of West Africa*, 2nd ed. (1968); WILLIAM WATSON, *Ancient Chinese Bronzes* (1962).

Decorative metalwork: LESLIE AITCHISON, *A History of Metals*, 2 vol. (1960); J. STARKIE GARDNER, *Ironwork* (various editions, 1892–1930); HERMANN LEISINGER, *Romanesque Bronzes: Church Portals in Mediaeval Europe* (1957); CYRIL STANLEY SMITH, *A History of Metallography: The Development of Ideas on the Structure of Metals Before 1890* (1960).

Enamelwork. *Materials and techniques:* KENNETH FRANCIS BATES, *Enameling: Principle and Practice* (1951); HERBERT MARYON, *Metalwork and Enamelling*, 4th ed. (1959), contains a useful bibliography; MARGARET SEELER, *The Art of Enameling* (1969).

Periods and centres of production: KLAUS WESSEL, *Byzantine Enamels from the 5th to the 13th Century* (1968); LUIGI MALLE, *Cloisonnés bizantini* (1970); SHALVA AMIRANASHVILI, *Medieval Georgian Enamels of Russia* (1964); MARY CHAMOT, *English Mediaeval Enamels* (1930); W.L. HILDBURGH, *Medieval Spanish Enamels and Their Relation to the Origin and the Development of Copper Champlevé Enamels of the Twelfth and Thirteenth Centuries* (1936); MARIE-MADELEINE GAUTHIER, *Emaux limousins champlevés des XIIe, XIIIe, and XIVe siècles* (1950); J.J. MARQUET DE VASSELOT, *Les Croises limousines du XIIe siècle* (1941); PAUL THOBY, *Les Croix limousines de la fin du XIIe siècle au debut du XIVe siècle* (1953); KATIA GUTH-DREYFUS, *Transluzides Email in der ersten Hälfte des 14 Jahrhunderts am Ober-, Mittel-, und Niederrhein* (1954); PHILIPPE VERDIER, *Catalogue of the Painted Enamels of the Renaissance in the Walters Art Gallery* (1967), with an excellent bibliography; HENRI CLOUZOT, *Dictionnaire des miniaturistes sur émail* (1924) and *La Miniature sur émail en France* (1928); PIERRE F. SCHWEDEBERGER, *Les Peintres sur émail genevois au XVIIe et au XVIIIe siècle* (1958); CHARLES BEARD, "Bavarian Enamels of the Seventeenth Century," *Connoisseur*, 97:267–271 (1936); EDWARD DILLON, "English Enamels on Brass of the Seventeenth Century," *Burlington Magazine*, 16:261 (1910); EGAN MEW, *Battersea Enamels* (1926), out of date in some respects; SANDOR MIHALIK, *Emailkunst im alten Ungarn* (1961; Eng. trans., *Old Hungarian Enamels*, 1961); HARRY GARNER, *Chinese and Japanese Cloisonné Enamels* (1962), with a full bibliography; LAWRENCE A. COBEN and DOROTHY C. FERSTER, *Japanese Cloisonne: History, Technique, and Appreciation* (1982). is meant for collectors and scholars.

Lacquerwork. *European lacquer:* HANS HUTH, *Lacquer of the West: The History of a Craft and an Industry, 1550–1950* (1971), a concise, well-written statement, the most complete to date, based on a life-time study, with many excellent illustra-

tions. ERICH KARSTEN, *Lackrohstoff-Tabellen*, 7th ed. (1981), a monograph mainly on tables; LUCY MAXYM, *Russian Lacquer, Legends and Fairy Tales* (1981); and OLGA VORONOVA (comp.), *Lacquer Miniatures from Mstiora* (1980), dealing with original folk miniature painting and the specific techniques developed in Central Russia.

Chinese lacquer: FILIPPO BONANNI, *Trattato sopra la vernice comunemente detta cinese* (1720); PERE L. D'INCARVILLE, "Mémoire sur le Vernis de la Chine," in the *Mémoires* of the Académie Royale des Sciences (1760); STEPHEN W. BUSHELL, *Chinese Art*, vol. 1 (1904); EDWARD F. STRANGE, *Catalogue of Chinese Lacquer in the Victoria and Albert Museum* (1925) and *Chinese Lacquer* (1926); R.S. JENYNS, "Chinese Lacquer," *Transactions of the Oriental Ceramic Society*, vol. 17 (1939–40); FRITZ LOW-BEER, "Chinese Lacquer of the Early 15th Century" and "Chinese Lacquer of the Middle and Late Ming Period," *Bulletin of the Museum of Far Eastern Antiquities*, no. 22 and 24 (1950–52); KURT HERBERTS, *Das Buch der ostasiatischen Lackkunst* (1959; Eng. trans., *Oriental Lacquer*, 1962); WERNER SPEISER, *Lackkunst in Ostasien* (1965); HARRY M. GARNER, "Diaper Backgrounds on Chinese Carved Lacquer," *Ars Orientalis*, 6:165–190 (1966), and "A Group of Chinese Lacquers with Basketry Panels," *Archives of Asian Art*, 20:6–24 (1966–67); LEE YU-KUAN, *Oriental Lacquer Art* (1972).

Japanese lacquer: JOHANN J. REIN, *Japan, nach Reisen und Studien*, vol. 2 (1886; Eng. trans., *The Industries of Japan*, vol. 2, 1889, reprinted 1969); MICHAEL TOMKINSON, *A Japanese Collection*, 2 vol. (1898); *L'Historie de l'art du Japon*, Paris Exhibition (1900); FRANK BRINKLEY, *Japanese Temples and Their Treasures*, 3 vol. (1910); *Official Catalogue of the Japan-British Exhibition* (1910); EDWARD F. STRANGE, "The Incense Ceremony and Its Utensils," Japan Society *Transactions*, 21:28–38 (1923–24), and *Catalogue of Japanese Lacquer and Inrō in the Victoria and Albert Museum* (1924); MARTHA BOYER, *Japanese Export Lacquers from the Seventeenth Century in the National Museum of Denmark* (1959); BEATRIX VON RAGUE, *Geschichte der japanischen Lackkunst* (1967). NATIONAL MUSEUM OF MODERN ART. TOKYO, *Japanese Lacquer Art: Modern Masterpieces* (1982; originally published in Japanese, 1981).

Mosaic. *General works:* A. BLANCHET, *La Mosaique* (1928); E.W. ANTHONY, *A History of Mosaics* (1935, reprinted 1968); P. FISCHER, *Das Mosaik* (1969). H.P. L'ORANGE and P.J. NORDHAGEN, *Mosaikk* (1958); Eng. trans., 1966), deals with the history of the art from its beginnings to *c.* AD 900.

Greek and Roman mosaic: Of the literature on Greek and Roman mosaic, the following works are of particular importance: (*pebble mosaic*): R.S. YOUNG, "Gordion 1956: Preliminary Report," *Am. J. Archaeology*, 61:319–331 (1957); D.M. ROBINSON, *Excavations at Olynthus*, vol. 5, 8, and 13 (1933, 1938, 1950). (*pebble-tessera mosaic*): C.M. ROBERTSON, "Greek Mosaics," *J. Hellenic Stud.*, 85:72–89 (1965); B.R. BROWN, *Ptolemaic Paintings and Mosaics and the Alexandrian Style* (1957); J. CHANDNARD, *Les Mosaïques de la maison des masques* (1933). (*floor mosaic—Roman*): *Colloque sur la mosaïque gréco-romaine, Actes* (1965); H.E. BLAKE, articles on Roman mosaics in *Mem. Am. Acad. Rome*, 8:7–160, 13: 69–214, and 17:81–130 (1930, 1936, 1940); K.M. PHILLIPS, "Subject and Technique in Hellenistic-Roman Mosaics: A Ganymede Mosaic from Sicily," *Art Bull.*, 42:243–262 (1960); R.P. HINKS, *Catalogue of the Greek, Etruscan and Roman Paintings and Mosaics in the British Museum* (1933); D. LEVI, *Antioch Mosaic Pavements* (1947); I. LAVIN, "The Hunting Mosaics of Antioch and Their Sources," *Dumbarton Oaks Papers*, no. 17 (1963). H. STERN, "Origine et débuts de la mosaïque murale," *Études d'archéologie classique*, vol. 2 (1959).

Early Christian mosaic: M. VAN BERCHEN and E. CLOUZOT, *Mosaïques chrétiennes du IVe au Xe siècle* 1924); G. MATTHIAE, *Mosaici medioevali delle chiese di Roma* (1967); W.F. OAKESHOTT, *The Mosaics of Rome, from the Third to the Fourteenth Centuries* (1967); H. STERN, "Les Mosaïques de l'Église de Sainte-Constance a Rome," *Dumbarton Oaks Papers*, no. 12 (1958); C. CECCHELLI, *I mosaici della basilica di S. Maria Maggiore* (1956); C. RICCI, *Tavole storiche dei mosaici di Ravenna*, 4 vol. (1930–37); F.W. DEICHMANN, *Frühchristliche Bauten und Mosaiken von Ravenna* (1958); G. BRUSIN and P.L. ZOVATTO, *Monumenti paleocristiani di Aquileia e di Grado* (1957); J.L. MAIER, *Le Baptistere de Naples et ses mosaïques* (1964); B. BRENK, "Die ersten Goldmosaiken der christlichen Kunst," *Palette* (Basel), 38:16–25 (1971), a work concerning the earliest use of gold cubes in mosaics.

Early Byzantine mosaic: G. BRETT *et al.*, *The Great Palace of the Byzantine Emperors* (1947); C. DIEHL, M. LE TOURMEAU, and H. SALADIN, *Les Monuments chrétiens de Salonique* (1918); T. SCHMIT, *Die Koimesis-Kirche von Nikaia* (1927); P.A. UNDERWOOD, "The Evidence of Restorations in the Sanctuary Mosaics of the Church of the Dormition at Nicaea," *Dumbarton Oaks Papers*, no. 13 (1959); M. AVIYONAH and M. SCHAPIRO, *Israel: Ancient Mosaics* (UNESCO 1954–60); E. KITZINGER, *Israeli Mosaics of the Byzantine Period* (1965); K. WEITZMANN, "The Mosaic in St. Catherine's Monastery on Mount Sinai," *Am. Phil. Soc. Proc.*, 110:392–405 (1966); K.A.C. CRESWELL, *Early Muslim Architecture*, 2nd ed., vol. 1 (1969), including K. VAN BERCHEK, "The Mosaics of the Dome of the Rock in Jerusalem and of the Great Mosque in Damascus"; G. BOVINI, *Mosaici di Ravenna* (1956; Eng. trans., *Ravenna Mosaics*, 1956), and *La vita di Cristo nei mosaici di S. Appollinare Nuevo di Ravenna* (1959). B.J. NORDHAGEN, "The Mosaics of John VII (705–707 AD)," *Acta. Institutum Romanum Nowegiae*, 2: 121–166 (1965); GUNILLA Å. KERSTRÖM-HOUGEN, *The Calendar and Hunting Mosaics of the Villa of the Falconer in Argos* (1974).

Middle and late Byzantine: DEMUS, *Byzantine Mosaic Decoration* (1948). A. GRABER, *La Peinture Byzantine* (1953; Eng. trans., *Byzantine Painting*, 1953); T. WHITTEMORE, *The Mosaics of Haghia Sophia at Istanbul*, 4 vol. (1933–52); P.A. UNDERWOOD and E.J.W. HAWKINS, "The Mosaics of Hagia Sophia at Istanbul: The Portrait of the Emperor Alexander," *Dumbarton Oaks Papers*, no. 15 (1961); E.J.W. HAWKINS, "Further Observations on the Marthex Mosaic in St. Sophia at Istanbul," *ibid.* 22 (1968); C.A. MANGO "Materials for the Study of the Mosaics of St. Sophia at Istanbul," *ibid.* 8 (1962); P.A. UNDERWOOD, *The Kariye Djami*, 3 vol. (1967); E. DIEZ and O. DEMUS, *Byzantine Mosaics in Greece, Hosios Lucas and Daphni* (1931); A. GRABAR and M. CHAIZIDAKIS, *Greece: Byzantine Mosaics* (UNESCO 1960); V.N. LAZAREV, *Old Russian Murals and Mosaics: From the 11th to 16th Century* (1966).

Medieval mosaics in the West: W.F. OAKESHOTT, *The Mosaics of Rome, from the Third to the Fourteenth Centuries* (1967); O. DEMUS, *The Mosaics of Norman Sicily* (1950); F. KITZENGER, "The Mosaics of the Cappella Palatina in Palermo," *Art Bull.*, 31:269–292 (1949); *The Mosaics of Monreale* (1960); A. DEWITT (ed.), *I mosaici del Battistera di Firenze*, 4 vol. (1954–59); H. KIER, *Der mittelalterliche Schmuckfussboden* (1970); A. PRANDI, "Pietro Cavallini a S. Maria in Trastevere," *Rivista dell'Istituto Nazionale d'Archeologia e Storia dell'Arte*, 1: 282–297 (1952); R. DERTEL, "Wandmalerei und Zeichnung in Italien," *Mitteilungen des Kunsthistorischen Institutes in Florenz*, 5:217–314 (1937–40), on the technique; F. FORLATI, "La tecnica dei primi mosaici marciani," *Arte veneta*, 3:85–87 (1949); HETTY JOYCE, *The Decoration of Walls, Ceilings, and Floors in Italy in the Second and Third Centuries A.D.* (1981).

Renaissance to modern: The best surveys are those contained in E.V. ANTHONY, *A History of Mosaics* (1935, reprinted 1968); and P. FISCHER, *Das Mosaik* (1969). Specific instances are discussed in FRANK R. DIFEDERICO, *The Mosaics of Saint Peter's: Decorating the New Basilica* (1983); and ALVAR GONZÁLEZ-PALACIOS, *The Art of Mosaics: Selections from the Gilbert Collection* (1982).

Stained glass. Illustrated monographs include: COMITE INTERNATIONAL D'HISTOIRE DE L'ART, *Corpus Vitrearum Medii Aevi* (1956–), a series intended when complete to document all medieval stained glass extant; J. BAKER and A. LAMMER, *English Stained Glass* (1960), probably the best collection of high-quality detail photographs of medieval stained glass ever published; H.E. READ, *English Stained Glass* (1926); A.C. SEWTER, *The Stained Glass of William Morris and His Circle* (1972); M. AUBERT *et al.*, *Le Vitrail Français* (1958), the standard work on French stained glass, although marred by many poor reproductions. This fault is remedied in part by the following works: M. AUBERT, *Stained Glass of the XIIth and XIIIth Centuries from French Cathedrals* (1947); and E. VON WITZLEBEN, *French Stained Glass* (1966). H. WENTZEL, *Meisterwerke der Glasmalerei*, 2nd ed. (1954), is the standard work on German medieval stained glass; E. VON WITZLEBEN, *Farbwunder Deutscher Glasmalerei aus dem Mittelalter* (1965) is the most copiously illustrated recent monograph on German medieval stained glass. S. BEEH-LUSTENBERGER, *Glasmalerei um 800–1900 im Hessischen Landesmuseum in Darmstadt* (1967), an excellent guide to one of the largest museum collections of medieval stained glass; P. WEMBER, *Johan Thorn Prikker: Glasfenster, Wandbilder, Ornamente 1891–1932* (1966); G. MARCHINI, *Italian Stained Glass Windows* (1957); M. STETTLER, *Swiss Stained Glass of the Fourteenth Century from the Church of Koenigsfelden* (1949); and F. ZSCHOKKE, *Medieval Stained Glass in Switzerland* (Eng. trans. 1947). See also JAMES STURM, *Stained Glass from Medieval Times to the Present: Treasures to Be Seen in New York* (1982); and ERNE R. FRUEH and FLORENCE FRUEH, *Chicago Stained Glass* (1983).

Books on aesthetic analysis include J.R. JOHNSON, *The Radiance of Chartres: Studies in the Early Stained Glass of the Cathedral* (1965), valuable not only for its somewhat too sweeping criticism of Viollet-le-Duc but for its original analysis of the effect of the twilight atmosphere of the cathedral upon our perception of its stained-glass windows; and R. SOWERS, *Stained Glass: An Architectural Art* (1965), a thoroughly illus-

trated analysis of the relation between stained glass and architecture, with numerous photographs of contemporary stained glass. DAVID EVANS, *A Bibliography of Stained Glass* (1982), is also recommended.

Stained-glass techniques are discussed in THEOPHILUS (RUGERUS), *Diversarum Artium Schedula* (1847; Eng. trans. 1963), the earliest account of stained-glass window-making techniques, now believed to have been written in the early 12th century; C. WINSTON, *Hints on Glass Painting* (1847), and his *Memoirs* (1865), two books containing some of the most thorough and perceptive analyses of medieval glass-painting techniques ever written; E. VIOLLET-LE-DUC, "Vitrail," in *Dictionnaire Raisonne de l'architecture française*, vol. 9 (1868; Eng. trans., "Mediaeval Stained Glass," 1946), still valuable for its many acute observations; and C.W. WHALL, *Stained Glass Work* (1905), a thorough craft manual by one of the leading turn-of-the-century stained-glass artists in England. Recent manuals include E.L. ARMITAGE, *Stained Glass* (1959); P. REYNTIENS, *The Technique of Stained Glass* (1967); and R. and G. MECALF, *Techniques of Stained Glass* (1971). Pictorial works can also be of interest: M.J. GRADL (ed.), *Authentic Art Nouveau Stained Glass Designs in Full Color* (1983); CONNIE EATON, *Oval Stained Glass Pattern Book* (1983); ANITA ISENBERG and SEYMOUR ISENBERG, *How to Work in Stained Glass*, 2nd ed. (1983).

History of glass design. There is ample literature on the history of glass. The following selection of titles includes basic reference works and handbooks as well as some specialized studies many of which contain bibliographical references. In addition, the *Journal of Glass Studies*, issued annually by The Corning Museum of Glass, includes extensive bibliographies.

The basic sources for medieval glassmanufacture are HERACLIUS, *Von den Farben und Künsten der Römer*, ed. by ALBERT ILG (1873); and THEOPHILUS PRESBYTER, *Schedula diversarum artium*, ed. by ALBERT ILG (1874; Eng. trans., *On Divers Arts: The Treatise of Theophilus*, 1963). GEORG AGRICOLA, *De re metallica* (1556; Eng. trans., 1912, reprinted 1950); and particularly ANTONIO NERI, *L'arte vetraria* (1612; Eng. trans. by CHRISTOPHER MERRET, *The Art of Glass . . .*, 1662), describe in detail glassmaking in the 16th and 17th centuries. See also JOHANN KUNCKEL, *Ars vetraria experimentalis*, 2 pt. (1679). Other technological studies are APSLEY PELLATT, *Curiosities of Glass Making* (1849); and ALFRED LUCAS, *Ancient Egyptian Materials and Industries*, 4th ed. rev. (1962). Development of glass technology in history is discussed in RUTH HURST VOSE, *Glass* (1980).

EDWARD DILLON, *Glass* (1907); ROBERT SCHMIDT, *Das Glas*, 2nd ed. (1922); and W.B. HONEY, *Glass: A Handbook . . . Victoria and Albert Museum* (1946), are among the best and most comprehensive general surveys of the history of glass. *Masterpieces of Glass* (1968), a catalog of some of the holdings in the British Museum, is the most recent scholarly publication on the subject in general, accompanied by a large bibliography. CHARLES G. JANNEAU, *Modern Glass* (1931), is a review of world glass at the beginning of the 1930s. For a general study of the international development of art glass, see ADA BUCH POLAK, *Modern Glass* (1962). GEOFFREY W. BEARD, *Modern Glass* (1968), provides a brief account of modern glasswork from various countries.

Comprehensive illustrative material on glass of the ancient world is found in GUSTAVUS A. EISEN and FAHIM KOUCHAKJI, *Glass*, 2 vol. (1927); the most scholarly survey is that of THE CORNING MUSEUM, *Glass from the Ancient World: The Ray Winfield Smith Collection* (1957). Roman glass in particular was treated exhaustively by ANTON KISA in *Das Glas im Altertume*, 3 vol. (1908). Basic treatises on pre-Roman glass include H.C. BECK, "Glass Before 1500 B.C.," in *Ancient Egypt and the East*, pt. 1, pp. 7–21 (June 1934); POUL FOSSING, *Glass Vessels Before Glass-Blowing* (1940); and BIRGIT NOLTE, *Die Glasgefässe im alten Ägypten* (1968). In addition to KISA (*op. cit.*), general books on Roman glass, such as MORIN-JEAN, *La Verrerie en Gaule sous l'Empire romain* (1913); CLASINA ISINGS, *Roman Glass from Dated Finds* (1957); and DONALD B. HARDEN, *Roman Glass from Karanis Found by the University of Michigan Archaeological Expedition in Egypt 1924–29* (1936), are important for the understanding of this period. The latter has become the standard reference work for describing and cataloging ancient glass in general.

Western glass of the 5th–8th centuries is treated in detail by D.B. HARDEN, "Glass Vessels in Britain and Ireland, A.D. 400–1000," in *Dark Age Britain* (1956). The standard handbooks on Islāmic and Western medieval glass are still CARL J. LAMM, *Mittelalterliche Gläser und Steinschnittarbeiten aus dem Nahen Osten*, 2 vol. (1929–30); and FRANZ RADEMACHER, *Die deutschen Gläser des Mittelalters* (1933). Byzantine glass is described in JOSEPH PHILIPPE, *Le Monde byzantin dans l'histoire de la verrerie, Ve–XVIe siècle* (1970). The basic handbooks on French and Belgian glass are JAMES BARRELET, *La Verrerie en France de l'époque gallo-romaine à nos jours* (1953); and RAYMOND CHAMBON, *L'Histoire de la verrerie en Belgique du IIe siècle à nos jours* (1955), the latter including ample bibliographic references on literary sources. WILLIAM A. THORPE, *A History of English and Irish Glass*, 2 vol. (1929), is still the standard reference work on English glass while HUGH WAKEFIELD COVERS *Nineteenth Century British Glass* (1961). German, Bohemian, and Austrian glass is treated exhaustively in ROBERT SCHMIDT, *Die Gläser der Sammler Mühsam*, 2 vol. (1914–27). For polychrome painting on vessels, see AXEL VON SALDERN, *German Enameled Glass* (1965). The handbooks on glass from *c.* 1800 to *c.* 1900 are GUSTAV PAZAUREK, *Gläser der Empire und Biedermeierzeit* (1923) and *Moderne Gläser* (1901). ASTONE GASPARETTO, *Il vetro di Murano dalle origini ad oggi* (1958), is the basic reference work on Venetian glass. The best surveys on Spanish glass are JOSEP GUDIOL Y RICART, *Los vidrios catalanes* (1941); and ALICE WILSON FROTHINGHAM, *Spanish Glass* (1964). For Scandinavian material, ADA BUCH POLAK, *Gammelt Norsk Glass* (1953); and HERIBERT SEITZ, *Äldre Svenska Glas . . .* (1936), should be consulted—both contain an English summary.

Glass in the United States has been dealt with in great detail by GEORGE S. and HELEN MCKEARIN in *American Glass* (1948) and *Two Hundred Years of American Blown Glass*, rev. ed. (1966). LURA W. WATKINS, *American Glass and Glassmaking* (1950, reprinted 1970), presents a useful outline of 19th- and 20th-century American glass. RAY and LEE GROVER, *Art Glass Nouveau* (1967), is valuable for its colour illustrations of 19th- and 20th-century fancy glasses in American collections. See also MARY JEAN MADIGAN, *Steuben Glass: An American Tradition in Crystal* (1982); and GERALD STEVENS, *Glass in Canada* (1982).

On Chinese glass, see W.C. WHITE, *Tombs of Old Lo-Yang* (1934); FRIEDRICH HIRTH, *China and the Roman Orient*, pp. 228–234 (1885); W.B. HONEY, "Early Chinese Glass," *Burlington Magazine*, 71:211–223 (1937); and H.C. BECK, "Far Eastern Glass: Some Western Origins," *Bulletin of the Museum of Far Eastern Antiquities*, 10:1–64 (1938).

Delhi

Delhi is the second largest city of India, surpassed in population only by Greater Bombay (Mumbai). New Delhi, the capital of India, lies immediately to the south of Delhi (popularly known as Old Delhi); both are within the Delhi national capital territory. Besides being at the political centre of the country, Delhi is also a focal point in India's transportation network.

Delhi is situated in north-central India about 100 miles (160 kilometers) south of the Himalayas and stands on the west (right) bank of the Yamuna River, a tributary of the Ganges. The national capital territory lies at an elevation between 700 and 1,000 feet (213 and 305 metres) and covers an area of 573 square miles (1,483 square kilometres). Of this area, Old Delhi occupies 360 square miles and New Delhi 169 square miles. The national capital territory is bounded to the east by the state of Uttar Pradesh and to the north, west, and south by Haryāna. It generally has been presumed that the city was named for Rājā Dhilu, a king of the 1st century BC, and that the various names by which it has been known (Delhi, Dehli, Dillī, and Dhillī) have been corruptions of this name.

Delhi has been the capital city of a succession of mighty empires and powerful kingdoms, and numerous ruins mark the sites of the various cities. According to popular tradition, the city has changed its locality a total of seven times, although some authorities, who take smaller towns and strongholds into account, claim it has changed its site as many as 15 times. All of these locations are confined to a triangular area of about 70 square miles called the Delhi triangle. Two sides of this triangle are represented by the rocky hills of the Arāvalli Range in the west and south and the third side by the shifting channel of the Yamuna River. The present site of Delhi is bounded to the west by a northern extension of the Arāvali Range known as the Delhi Ridge.

This article is divided into the following sections:

Physical and human geography 221
- The landscape 221
  - Climate
  - Plant and animal life
  - The city layout
- The people 224
- The economy 224
  - Industry
  - Finance and trade
  - Transportation
- Administration and social conditions 224
  - Government
  - Housing
  - Public utilities
  - Health and security
  - Education
- Cultural life 225

History 225

Bibliography 226

Physical and human geography

THE LANDSCAPE

Climate. The climate of Delhi is characterized by extreme dryness, with intensely hot summers. It is associated with a general prevalence of continental air, which moves in from the west or northwest, except during the season of the monsoon (rain-bearing wind), when an easterly to southeasterly influx of oceanic air brings increased humidity. The summer season lasts from mid-March to the end of June, with average maximum and minimum temperatures of 97° F (36° C) and 77° F (25° C); it is characterized by frequent thunderstorms and squalls, which are most frequent in April and May. The monsoon season, following the hot summer, continues until the end of September, with an average rainfall of about 26 inches (660 millimetres). The post-monsoon period of October and November constitutes a transition period from monsoon to winter conditions. The winter season extends from late November to mid-February. The air in Delhi is dry for most of the year, with low relative humidity from April to June and markedly higher humidity in July and August, when weather conditions are oppressive. Delhi's mean daily temperature is highest in May; and the monthly mean temperature is highest in June, which is also the month when the night temperature is at its maximum. The mean daily temperature may rise as high as 110° F (43° C). The coldest month is January, when both the mean maximum temperature and the mean minimum temperature are at their lowest—70° F (21° C) and 45° F (7° C), respectively.

Temperature and rainfall

Air and water pollution have increased with the growth of population, industry, and the use of motor vehicles. Sometimes a temperature inversion (which can occur when a warm air mass remains over a land surface that cools during the night) forms in the winter months, which traps pollutants, prevents them from dispersing, and increases contamination considerably.

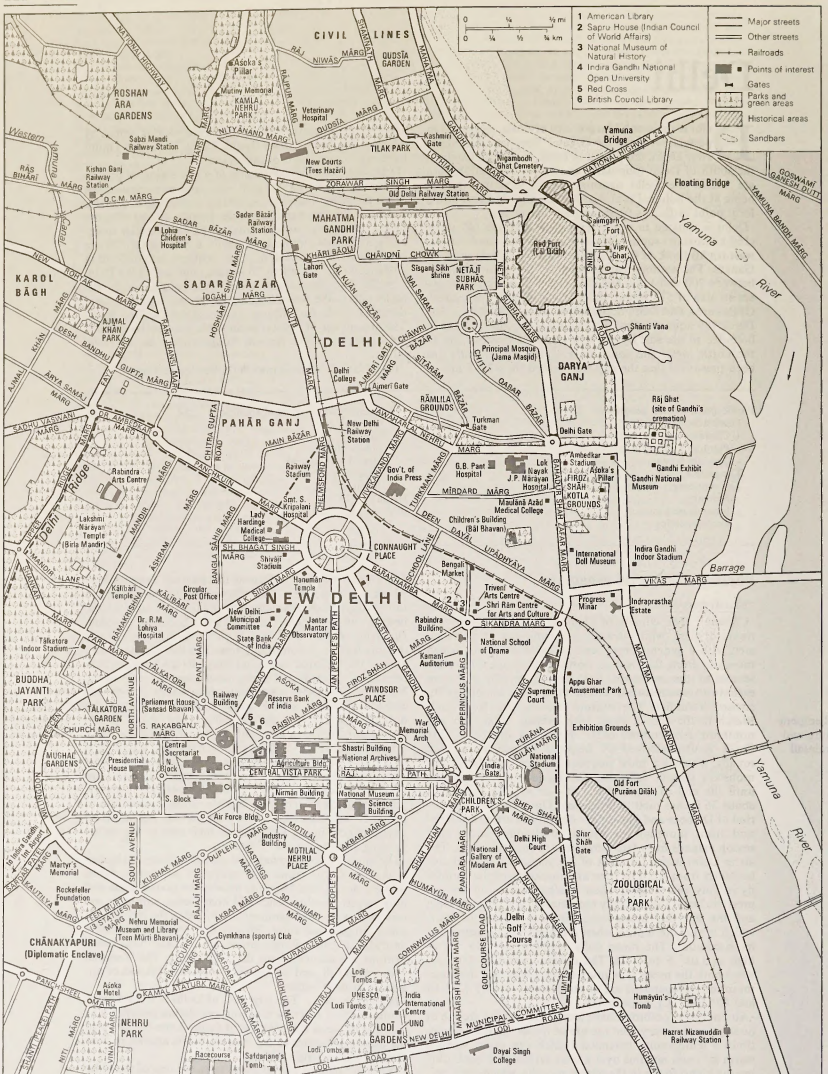
Plant and animal life. The natural plant cover in the Delhi area varies according to the physical features with which it is associated. The ridges and hillsides abound in thorny trees, such as acacias. During the monsoon season, herbaceous species grow in profusion. The sissoo (shisham; *Dalbergia sissoo*) tree, which yields a dark brown and durable timber, is commonly found in the Bāngar (Plain) area of the national capital territory. Riverine vegetation, consisting of weeds and grass, occurs on the banks of the Yamuna. New Delhi is known for its avenues of flowering shade trees, such as the neem (*Azadirachta indica;* a drought-resistant tree with a pale yellow fruit), jaman (*Syzygium cumini;* a tree with an edible grapelike fruit), mango, pipal (*Ficus religiosa;* a fig tree), and sissoo. It is also known for numerous flowering plants, which provide a splash of colour during the winter. These include a large number of multicoloured seasonals: chrysanthemums, phlox, violas, and verbenas. The transition from winter to spring is very gradual, and only the flowers can testify to changing conditions, with chrysanthemums in December yielding place to roses in February.

New Delhi's avenues of shade trees

The animal life of the national capital territory is similarly diverse. Carnivorous mammals such as hyenas, foxes, wolves, jackals, and leopards inhabit the ravine lands and hilly ridges. Antelope and wild boars, once common, are now rare in the wild. Monkeys are found in the city around some temples and ruins. Birdlife is profuse and includes partridge, pigeons, jungle crows, parrots, and bush quail. Peafowl are numerous on the hilly ridges. The Yamuna abounds in fish, and the region's lakes attract flocks of migratory birds in winter.

The city layout. The city plan of Delhi is a mixture of contrasting old and new road and circulation patterns. The contrast between the convoluted form of the old city and the diagonal features of the modern traffic arteries in New Delhi is particularly striking.

The street pattern of Old Delhi reflects some of the older requirements of defense, with a few transverse streets leading from one major gate to another. Occasionally a through street from a subsidiary gate leads to the main



Central Delhi–New Delhi.

Based upon Survey of India map with permission of the Surveyor General of India, © Government of India Copyright 1969

axes. The other Old Delhi streets tend to be irregular in direction, length, and width and are suitable only for pedestrian traffic. Thus, the pattern as a whole consists of a confusing mixture of narrow and winding streets, culs-de-sac, alleys, and byways giving access to residences and commercial areas.

In sharp contrast to Old Delhi, the Civil Lines (residential areas originally built by the British for senior officers) in the north and New Delhi in the south present an aspect of relative openness, characterized by green grass and trees, order, and quiet.

When the decision was made in 1911 to transfer the capital of India from Calcutta to Delhi, and a town planning committee was formed, a site was chosen three miles south of the existing city of Delhi, around Rāisīna Hill. This was a well-drained, healthy area between the ridge and the river that provided ample room for expansion. The Rāisina Hill, commanding a view of the entire area, stood 50 feet (15 metres) above the plain, but the top 20 feet were blasted off to make a level plateau for the major government buildings and to fill in depressions. With this low acropolis as the focus, the plan was laid out.

The New Delhi plan was characterized by wide avenues with trees in double rows on either side, creating vistas and connecting various points of interest. Almost every major road has a specific focal point closing the vista so that no avenue is lost in the horizon. Besides the diagonal road pattern, the most prominent feature of the plan is the Central Vista Park, starting from the National Stadium in the east, continuing through the All India War Memorial Arch (popularly called the India Gate) and the Central Secretariat (Kendriya Sachivālaya), and culminating in the west at the Presidential House (Rāshtrapati Bhavan). This is the main east–west axis; it divides New Delhi into two parts, with the fashionable shopping centre, Connaught Place, in the north and extensive residential colonies in the south.

The Central Vista Park

Land use. The pattern of land use in Delhi was influenced considerably by the implementation (albeit partial) of the Delhi Development Authority's 20-year (1962–81) master plan. Broadly, public and semipublic land use was concentrated in the Central Secretariat area of New Delhi and in the Old Secretariat area in the Civil Lines, with subsidiary centres developing in the Indraprastha Estate (an office complex) in the east and in Ramakrishnapuram (an office-cum-residence complex) in the south. A large number of small manufacturing establishments have entrenched themselves in almost every part of Old Delhi, but the main industrial areas have become concentrated along Najafgarh Mārg (Shrivāji Mārg), in the west, and on Mathura Mārg, in the south, where a large planned industrial estate (Okhla) has been established. Areas for commercial land use are confined mainly to Chāndnī Chowk and Khāri Bāoli (both in the north), the Sadar Bāzār of Old Delhi, the Ajmal Khān Mārg of Karol Bāgh in western Delhi, and the Connaught Place area of New Delhi. A number of district and local shopping centres have also developed in other localities.

The University of Delhi is located in the north, where a number of educational institutions for college education and for higher studies are located. Another educational complex that includes Jawaharlal Nehru University, the Indian Institute of Technology, and other institutions has been developed in southern Delhi.

Traditional areas. In a city such as Delhi, which bears the impress of history, there is a clear distinction between areas where indigenous influences are uppermost and areas characterized by colonial and modernizing influences. Although the social structure of Delhi has changed from coherence to a heterogeneity that is in keeping with its position as the national capital, certain residential neighbourhoods in Old Delhi, in the Civil Lines, in government housing areas, and in more recently developed areas have acquired a specific character of their own.

In Old Delhi there is a strong feeling of *mohalla* ("neighbourhood"), partly induced by the peculiar housing layout. There gates or doorways open onto private residences and courtyards or onto *katrā* (one-room tenements facing a courtyard or other enclosure and having access to the street by only one opening or gate). The Civil Lines area consists of residences for upper income groups. The government housing areas also exhibit segregation by income groups. In some developed areas, "mixed neighbourhoods" have been created. Chānakyapuri (more commonly known as the Diplomatic Enclave), with its concentration of foreign embassies, represents a microcosm of international architecture. Cultural "islands" have formed in such areas as the Bengali Market area or Karol Bāgh; the latter, for example, is characterized variously by Bengali, South Indian, and Punjabi cultures, although cultural distinctiveness is being eroded as other city residents move in. Another facet of the city profile is the slum, inhabited mostly by construction workers, sweepers, factory labourers, and other low-income groups. There are also urban village enclaves, such as Kotla Mubārakpur, where houses and streets retain rural characteristics though residents have urban occupations.

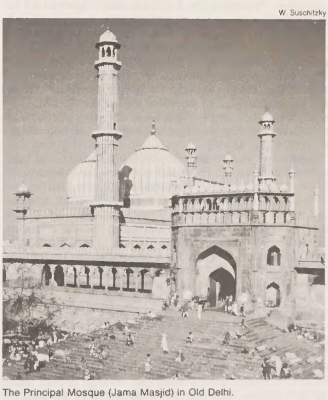
Architecture. There is perhaps no city in India that can compare with Delhi in the number of its monuments. These edifices illustrate the types of Indian architecture from the time of the imperial Gupta dynasty 1,600 years ago to the period of British rule, when the style of such architects as Sir Edwin Lutyens and Sir Herbert Baker was in evidence in New Delhi. Delhi is particularly rich in material for the study of Indo-Muslim architecture. The monuments of the early Pashtun style (1193–1320)—represented by the Qūwat-ul-Islām mosque, the Quṭb Minar, the tomb of Iltutmish, and the 'Alā'ī Gate—reveal the adoption and adaptation of Hindu materials and style to Islāmic motifs and requirements. The later Pashtun styles represented in Tughlakābād and in the tombs of the Sayyid kings (1414–51) and Lodī kings (1451–1526) are characterized by finer domes and decoration and the use of finer marbles and tiles. The later Mughal architecture represented in the Red Fort (Lāl Qilāh) and the Principal Mosque (Jama Masjid) reveals an increasing use of marble, elaboration of external surfaces with florid decoration, and the construction of bulbous domes and lofty minarets.

Delhi's monuments

The Red Fort is one of the most important buildings of the city. Its massive red sandstone walls, 75 feet in height, enclose a complex of palaces, gardens, military barracks, and other buildings. The two most famous of these are the Hall of Public Audience (Divān-e 'Āmm) and the Hall of Private Audience (Divān-e Khāṣṣ). The Hall of Public Audience has 60 red sandstone pillars supporting a flat roof. The Hall of Private Audience is smaller and has a pavilion of white marble.

The architectural styles in the British period—represented by the Central Secretariat, Parliament House

W. Suschitzky



The Principal Mosque (Jama Masjid) in Old Delhi.

(Sansad Bhavan), and the Presidential House (formerly the British viceroy's house)—combine the best features of the modern English school of architecture with traditional Indian forms. In the postindependence era, public buildings in Delhi began to show a utilitarian bias and a search for a synthesis of Indian and Western styles; the attempt, however, has not always been successful, as is evident from the Supreme Court building, the Science Building (a conference hall), and the government ministries. The Children's Building (a children's centre) and Rabindra Building (a fine arts centre) show a trend toward a new style, using modern materials. Along the Yamuna riverfront, memorials set in flowering gardens have been built for such 20th-century national leaders as Mahatma Gandhi (Rāj Ghat), Jawaharlal Nehru (Shānti Vana), and Lal Bahadur Shastri (Vijay Ghat).

THE PEOPLE

The refugee immigration

Delhi's population has increased dramatically from the 240,000 inhabitants it had in 1911. The highest growth rate occurred between 1941 and 1951—with the influx of refugees into the city at the time of independence—and population has also risen rapidly since 1990. Much of the increase is still from immigration. Population densities are among the highest of India's states and territories.

The composition of Delhi's population reflects its truly cosmopolitan character, with more than half of the residents coming from outside the territory. Most of these immigrants come from other Indian states and adjacent countries, and only a small proportion consists of resident foreigners. The religious composition of the population is also varied. The great majority of the population is Hindu; Muslims constitute the largest minority, followed by smaller numbers of Sikhs, Jains, Christians, and Buddhists.

THE ECONOMY

Delhi's service sector is the most important component of its economy and is the largest employer. The industrial sector is second and the commercial sector third. Agriculture once contributed significantly to the economy of the national capital territory but now is of little importance. A substantial proportion of Delhi's working population is engaged in various services, including public administration, the professions, the liberal arts, and various personal, domestic, and unskilled-labour services. As a trading and commercial centre, Delhi has held a dominant position in northern India for many centuries. In modern times it has also become a manufacturing centre and one of India's most important sources of export goods.

Industry. Traditionally, Delhi has been renowned for its artistic work, such as ivory carving and painting, gold and silver embroidery, decorative ware, copperware, and brassware. In modern times industry has diversified, and Delhi has become important for the manufacture of sophisticated products in small-scale industry, such as electronics and engineering goods, automobile parts, precision instruments, machinery, and electrical appliances. Wearing apparel, sports and leather goods, handloom products, and handicrafts are also produced. A large and thriving tourist sector has also developed.

Finance and trade. Delhi's position as the national capital and as a major industrial city have accentuated its function as a banking, wholesale-trade, and distribution centre. It is the headquarters of the Reserve Bank of India and of the regional offices of the State Bank and other banking institutions. It is also a divisional headquarters for the insurance business and an important stock-exchange centre. Delhi has long acted as a major distribution hub for much of northern India, handling a wide variety of items. Much of the distributive trade is carried on from within the Old Delhi area, where most of the markets are located near each other.

Transportation. The geographic position of Delhi on the great plain of India, where the Deccan tableland and the Thar Desert approach the Himalayas to produce a narrow corridor, ensures that all land routes from northwestern India to the eastern plain must pass through it, thus making it a pivotal centre in the subcontinent's network of transportation. Five national highways converge on Delhi. Several railway lines also meet there, linking the city with all parts of the country. Delhi is the most important air terminus in northern India for both domestic and international air services. Indira Gandhi International Airport, located in the southwestern part of the city, handles international flights. The nearby Pālam Airport is one of the hubs of the domestic airway system.

The traffic-circulation pattern within a city that was designed for a smaller population became heavily overburdened with Delhi's explosive growth. Improvements to the road system—such as adding overpasses and underpasses and widening major thoroughfares—have alleviated the worst traffic congestion, but the sheer volume of traffic—which includes such slow-moving vehicles as bullock carts, pedicabs, and bicycles—makes road travel in Delhi difficult, particularly during peak-hour conditions. Mass-transportation facilities are inadequate, the principal means of public transport consisting of buses. Long-distance commuting within the city is facilitated by Ring Road bus service and by the Ring Railway. The first sections of a planned 125-mile municipal rapid-transit system opened in 2002 and 2003.

ADMINISTRATION AND SOCIAL CONDITIONS

The administration

Government. Delhi was a chief commissioner's province when India attained independence in 1947. It became a centrally administered state in 1952, but in 1956 its status was changed to that of a union territory under the central government. Its designation was changed to the national capital territory in the early 1990s. A unified corporation for both urban and rural areas was established in 1958. The administrative system was further modified by the Delhi Administration Act of 1966. Under the present arrangement, Delhi has a three-tier administration consisting of a lieutenant governor and an executive council, an elected metropolitan council, and the municipal corporation. The lieutenant governor, appointed by the president of India, is the chief administrator of the territory and is assisted by an executive council of four members (headed by a chief executive councillor), who are also appointed by the president. The metropolitan council is a purely deliberative body. The municipal corporation is an elected local body having under its control most statutory autonomous bodies, notable exceptions being the New Delhi Municipal Committee, the Delhi Cantonment Board, and the Delhi Development Authority. The New Delhi Municipal Committee is a body nominated by the central government. The Delhi Cantonment Board consists of some elected and some nominated members, the latter with some ex officio members.

Housing. The housing situation in Delhi deteriorated after 1947 as a result of the influx of refugees caused by the partition of India and Pakistan and the city's emergence as the national capital of India. Since then, building activity has been insufficient to close the gap or to keep pace with the increasing population. This has compelled a large proportion of the city's population to seek shelter in congested areas and in unauthorized dwellings or to settle as squatters in slums.

The traditional houses in Old Delhi are unplanned, consisting of old structures of two, three, or more stories with a high proportion of single-room dwelling units. In the Civil Lines area there are a number of old one-story bungalows. In New Delhi the government housing colonies have been laid out in a lavish manner and are grouped by income.

A program to build new and rehabilitate old housing has been pushed since the 1950s; it is administered by a number of agencies, such as the government of the national capital territory, the various municipal governments, the Delhi Development Authority, and various individuals and cooperatives.

Public utilities. Water supply, drainage, sewerage, and conservancy and scavenging services are mandated functions of the municipal corporation. Such functions as city transportation and the generation and distribution of electricity, though not obligatory, are performed by the corporation. Three statutory agencies—the Delhi Water Supply and Sewage Disposal Undertaking, the Delhi Electric Sup-



The Central Secretariat in New Delhi.
Hubertus Kanus—Rapho/Photo Researchers

ply Undertaking, and the Delhi Transport Corporation—perform these functions.

Water supply

The supply of drinking water in Delhi has not kept up with demand, in spite of the fact that the water system has been improved and augmented several times. The Yamuna River, the main source of supply, is practically dry during the summer months. Underground water in the territory has generally been found to be brackish; Delhi, therefore, must depend for part of its needs upon the adjoining states.

Most of the residents of Delhi do not have access to adequate sewage disposal. Improvement is needed, both by way of extension of sewerage to new areas and by the expansion of its capacity in older areas. The treatment of sewage is also inadequate.

Delhi's electric power supply depends on power generated by local coal-burning thermal stations, augmented by sources outside the national capital territory. As with other utilities, the supply of power has always been distributed disproportionately.

Health and security. Overall health standards in Delhi exceed the national average, but the accessibility of health-care facilities varies widely. Much of the city's health care is provided by a large number of allopathic dispensaries, Āyurvedic and Unanī (*yūnānī*) dispensaries (*i.e.*, practicing indigenous systems of medicine that use mostly herbs and minerals), and homeopathic dispensaries. Most of the larger hospitals—such as the Dr. Ram Manohar Lohia Hospital, Smt. Sucheta Kripalani Hospital, and Lok Nayak J.P. Nārāyan Hospital—are administered by the national government or the Delhi administration.

The jurisdiction of the Delhi Fire Service extends over both the urban and rural areas of the national capital territory. In the rural areas, temporary stations are opened during the summer. The Delhi Police Service is under a commissioner of police of the Delhi administration. The city is divided into four police districts, each of which is under a superintendent of police.

Education. The growth of modern education in Delhi has kept pace with the expansion of the city's population. Primary-level education is nearly universal, and a large proportion of students also attend secondary school. Education for women at all levels has advanced at a much faster pace than it has for men. Among the institutions of higher learning, the most important is the University of Delhi, which has many affiliated colleges and research institutions. Among the major colleges for professional and other studies are the Indian Agricultural Research Institute, the Indian Institute of Technology, and the All India Institute of Medical Sciences.

CULTURAL LIFE

Delhi's cultural life has been influenced considerably by the cosmopolitan character of its population, which comes from different parts of India and the world and possesses varied cultural backgrounds. Much has been borrowed and adapted from Western culture, a process accelerated since independence by the influence of the modern mass media. Television, however, has also facilitated a greater awareness of regional and national interests. Although the cultural activities of earlier days—such as dancing, music, and poetry forums (*mushā'ira*)—have been yielding place to the cinema, the cabaret, and clubs, there are also theatre groups and institutions that have fostered indigenous literature and fine arts. Many of India's major cultural institutions—including the national academies of music, dance, and drama; of art; and of letters—are located in Delhi, as are numerous libraries, archives, and museums.

Delhi is home to numerous fairs and festivals. In addition to a variety of trade and book fairs, the city hosts an annual film festival. The many religious groups in Delhi contribute to an ongoing succession of religious festivals and celebrations.

Gardens

Delhi is a city of gardens and fountains, notable examples being the Roshan Āra Gardens and the meticulously planned and laid out Mughal Gardens. Many park and garden areas have grown up around historical monuments, such as the Lodī Gardens (around the Lodī Tombs) and the Firoz Shāh Kotla Grounds (around Aśoka's Pillar). Among the major recreation areas are the Delhi Ridge and the Yamuna riverfront. Delhi has well-developed sporting facilities, many of which were built when the city hosted the Asian Games in 1982.

History

The earliest reference to a settlement at Delhi is found in the epic *Mahābhārata* (a narrative about the descendants of the prince Bharata), which mentions a city called Indraprastha, built about 1400 BC under the direction of Yudhiṣṭhira, a Pāṇḍava king, on a huge mound somewhere between the sites where the historic Old Fort (Purāna Qilāh) and Humāyūn's Tomb were later to be located. Although nothing remains of Indraprastha, according to legend it was a thriving city. The first reference to the place-name Delhi, as already mentioned, seems to have

been made in the 1st century BC, when Rājā Dhilu built a city near the site of the future Quṭb Minar and named it for himself. Thereafter Delhi faced many vicissitudes and did not reemerge into prominence until the 12th century AD, when it became the capital of the Cauhān (Cāhamāna) ruler Pṛthvīrāja III. After the defeat of Pṛthvīrāja in the late 12th century, the city passed into Muslim hands. Quṭb-ud-Dīn Aybak, founder of the Mu'izzī (Slave) dynasty and builder of the famous tower Quṭb Minar (completed in the early 13th century), also chose Delhi as his capital.

'Alā'-ud-Dīn Khaljī (1296–1316) built the second city of Delhi at Siri, three miles northeast of the Quṭb Minar. The third city of Delhi was built by Ghiyāṣ-ud-Dīn Tughluq (1320–25) at Tughlakābād but had to be abandoned in favour of the old site near the Quṭb Minar because of a scarcity of water. His successor, Muḥammad ibn Tughluq, extended the city farther northeast and built new fortifications around it. It then became the fourth city of Delhi, under the name Jahānpanāh. These new settlements were located between the old cities near the Quṭb Minar and Siri Fort. Muḥammad ibn Tughluq's successor, Fīrūz Shāh Tughluq, abandoned this site altogether and in 1354 moved his capital farther north near the ancient site of Indraprastha and founded the fifth city of Delhi, Fīrūzābād, which was situated in what is now the Firoz Shāh Kotla area.

Invasion of Timur

After the invasion and sack of Delhi by Timur (Tamerlane) at the end of the 14th century, the last of the sultan kings moved the capital to Āgra, so that Delhi experienced a temporary diminution in its importance. Bābur, the first Mughal ruler, reestablished Delhi as the seat of his empire in 1526. His son Humāyūn built a new city, Dīn Panāh, on the site between Firoz Shāh Kotla and the Purāna Qal'ah. Shēr Shāh, who overthrew Humāyūn in 1540, razed Dīn Panāh to the ground and built his new capital, the Shēr Shāhī (Purāna Qal'ah), as the sixth city of Delhi.

Delhi later again lost importance when the Mughal emperors Akbar (1556–1605) and Jahāngīr (1605–27) moved their headquarters, respectively, to Fatehpur Sīkri and Āgra, but the city was restored to its former glory and prestige in 1638, when Shāh Jahān, Akbar's grandson, laid the foundations of the seventh city of Delhi, Shāhjahānābād, which has come to be known as Old Delhi. The greater part of the city is still confined within the space of Shāh Jahān's walls, and several gates built during his rule—the Kashmirī Gate, the Delhi Gate, the Turkman Gate, and the Ajmerī Gate—still stand.

With the fall of the Mughal Empire during the mid-18th century, Delhi again faced many vicissitudes—raids by the Marāthā (a people of peninsular India), the invasion by Nāder Shāh of Persia, and a brief spell of Marāthā rule—before the arrival of the British in 1803. Under British rule the city flourished, except during the Indian Mutiny in 1857, when the mutineers seized the city for several months, after which British power was restored and Mughal rule ended. In 1912 the British moved the capital of British India from Calcutta to the partially completed New Delhi, the construction of which was finished by 1931. (V.L.S.P.R./K.V.Su./V.R.)

The modern city

Since India's independence, Delhi has grown far beyond its original boundaries, spreading north and south along the Yamuna River, spilling onto the river's east bank, expanding over the Delhi ridge to the west, and, eventually, extending beyond the boundaries of the union territory into adjacent states. This increase was initially in response to the huge influx of Hindu refugees from Pakistan following partition, but since the early 1950s Delhi began absorbing immigrants from throughout India at an astounding rate. New Delhi, once adjacent to Delhi, is now a part of the larger city, as are the sites of the former seats of empire. Between ancient mausoleums and forts have sprouted high-rise towers, commercial complexes, and other aspects of the modern city.

This rapid development has not been without cost, however. In a pattern familiar to many postcolonial megalopolises, the huge influx of job-seeking immigrants placed a colossal strain on the city's infrastructure and on the ingenuity of city planners to provide sufficient power, sanitation, and clean water for the population. Most problematic, in a city in which the population had more than doubled since 1980, fully one-tenth of Delhi's residents lived in urban slums called *jhuggi-jhompris*; these lacked the most basic services and left city planners and administrators the difficult task of integrating more than a million slum-dwelling residents into a city whose infrastructure failed to accommodate already existing households.

Further, traffic congestion in Delhi has become among the worst in the world, a situation that contributed greatly to the city's already hazardous level of air pollution—this earned the Indian capital the dubious honour of being among the most polluted cities in the world. Antipollution measures undertaken since the 1980s have improved Delhi's air quality considerably, but overcrowding, congestion, and an overburdened infrastructure have remained as major obstacles for the city to overcome. (Ed.)

BIBLIOGRAPHY. Descriptive works, with maps and illustrations, include P.R. MEHENDIRATTA, *Coming to India and to Delhi* (1972); INDIA TOURISM DEVELOPMENT CORPORATION, *Guide to Delhi*, 2nd ed. (1982); KHUSHWANT SINGH, *Delhi: A Portrait* (1983); and RICHARD PLUNKETT and HUGH FINLAY, *Delhi*, 2nd ed. (2000).

Analyses of social and economic conditions and ethnic developments in the Delhi metropolitan area are given in V.K.R.V. RAO and P.B. DESAI, *Greater Delhi: A Study in Urbanisation, 1940–1957* (1965); UNITED NATIONS DEPT. OF INTERNATIONAL ECONOMIC AND SOCIAL AFFAIRS, *Population Growth and Policies in Mega-Cities: Delhi* (1986); BISWAJIT BANERJEE, *Rural to Urban Migration and the Urban Labour Market: A Case Study of Delhi* (1986); and ASHOK RANJAN BASU, *Urban Squatter Housing in Third World* (1988).

For historical accounts, see GORDON RISLEY HEARN, *The Seven Cities of Delhi*, 2nd ed. (1928); PRABHA CHOPRA (ed.), *Delhi, History and Places of Interest*, rev. ed. (1975); NARAYANI GUPTA, *Delhi Between Two Empires, 1803–1931: Society, Government, and Urban Growth* (1981); H.K. KAUL (ed.), *Historic Delhi: An Anthology* (1985, reprinted 1996); and R.E. FRYKENBERG (ed.), *Delhi Through the Ages: Essays in Urban History, Culture, and Society* (1986). (V.R./Ed.)

Democracy

The term *democracy* literally means rule by the people. It is derived from the Greek *dēmokratiā,* which was coined from *dēmos* ("people") and *kratos* ("rule") in the middle of the 5th century BC to denote the political systems then existing in some Greek city-states, notably Athens.

The study of the history of democracy must consider two closely related topics. It is first of all concerned with the origins and development of democratic government, including the various forms such governments have taken throughout the world since the time of the ancient Greeks. It also encompasses the history of democratic ideas, including ideas about democracy's ultimate nature and value. This article will treat the history of democracy from both of these perspectives.

For coverage of related topics in the *Macropædia* and *Micropædia,* see the *Propædia,* sections 541, 542, 543, 912, 10/51, and 10/52, and separate articles on individual political philosophers (*e.g.,* ARISTOTLE and LOCKE).

This article is divided into the following sections:

Fundamental questions 226A
Democratic institutions 226A
  Prehistoric forms of democracy
  Classical Greece
  The Roman Republic
  The Italian republics from the 12th century to the Renaissance
  Toward representative democracy: Europe and North America to the 19th century
  The spread of democracy in the 20th century
  Contemporary democratic systems
The theory of democracy 226E
  Democratic ideas from Pericles to Rawls
  The value of democracy
Problems and challenges 226H
  Inequality of resources
  Immigration
  Terrorism
  International systems
  Transition, consolidation, breakdown
Bibliography 226H

Fundamental questions

If a government of or by the people is to be established, at least five fundamental questions must be confronted at the outset, and two more are almost certain to be posed if the democracy continues to exist for long.

(1) What is the appropriate unit or association within which a democratic government should be established? A town or city? A country? A business corporation? A university? An international organization? All of these?

(2) Who among the members of the association should enjoy full citizenship? Which persons, in other words, should constitute the *dēmos*? Should the *dēmos* include all adults? If only a subset of adults are included, how small can the subset be before the association ceases to be a democracy and becomes something else, such as an aristocracy (government by the best, *aristos*) or an oligarchy (government by the few, *oligos*)?

The *dēmos*

(3) How are citizens to govern? What political organizations and institutions will they need?

(4) When citizens are divided on an issue, as they often will be, whose views should prevail, and in what circumstances? Should a majority always prevail, or should minorities sometimes be empowered to block or overcome majority rule?

(5) If a majority is ordinarily to prevail, what is to constitute a proper majority? A majority of all citizens? A majority of voters? Should a proper majority comprise not individual citizens but certain groups or associations of citizens, such as hereditary groups or territorial associations?

(6) Why should "the people" rule? Is democracy really better than aristocracy or monarchy? Perhaps, as Plato argues in the *Republic,* the best government would be led by a minority of the most highly qualified persons—an aristocracy of "philosopher-kings." What reasons could be given to show that Plato's view is wrong?

(7) Finally, what conditions favour the continued existence of democracy? What conditions are harmful to it? Why have some democracies managed to endure, even through periods of severe crisis, while so many others have collapsed?

The Granger Collection, New York



Queen Elizabeth I presiding over Parliament, engraving, 1608.

Democratic institutions

PREHISTORIC FORMS OF DEMOCRACY

Although it is often asserted that democracy was created in Greece about the year 500 BC, it is very likely that democratic government, in a broad sense, existed in several areas of the world well before this time. Forms of direct democracy probably existed among many tribal groups during the thousands of years when human beings survived by hunting and gathering. When humans began to settle in fixed communities for agriculture and trade, the conditions that favour popular participation in government seem to have become rare. Greater inequalities in wealth and military power between communities, together with a marked increase in the typical community's size, encouraged the spread of hierarchical and authoritarian forms of social organization. Popular governments among settled peoples vanished, to be replaced for thousands of years by monarchies, despotisms, aristocracies, and oligarchies.

CLASSICAL GREECE

During the Classical period (corresponding roughly to the 5th and 4th centuries BC), Greece was of course not a country in the modern sense but a collection of several hundred independent city-states, each with its surrounding countryside. In 507 the citizens of Athens began to develop a system of popular rule that would last nearly two centuries. Thus the Athenian answer to question 1 was that the political association most appropriate to democratic government is the polis, or city-state.

The Athenian answer to question 2—Who should constitute the *dēmos*—was similar to the answer developed in many newly democratic countries in the 19th and 20th centuries. Although citizenship in Athens was hereditary, extending to anyone who was born to parents who were

themselves Athenian citizens, membership in the *dēmos* was limited to male citizens 18 years of age or older (until 403, when the minimum age was raised to 20). One scholar has suggested that in the mid-4th century there were about 100,000 citizens, 10,000 resident foreigners, or metics, and as many as 150,000 slaves. Among citizens, about 30,000 were males over 18. If these numbers are roughly correct, then the *dēmos* comprised 10 to 15 percent of the total population.

Regarding question 3—What political institutions are necessary for governing?—the Athenians adopted an answer that would appear independently elsewhere. The heart and centre of their government was the Assembly (Ecclesia), which met on the Pnyx, a hill west of the Acropolis. Decisions were taken by vote, and, as in many later assemblies, voting was by a show of hands. The votes of a majority of those present and voting prevailed. The agenda of the Assembly was set by the Council of Five Hundred, which, unlike the Assembly, was composed of representatives chosen by lot from each of 139 small territorial entities, known as demes. The number of representatives from each deme was roughly proportional to its population.

The Athenian Assembly

Another important political institution in Athens was the popular courts (*dikasteria*), described by one scholar as having "unlimited power to control the Assembly, the Council, the magistrates, and political leaders." The popular courts were composed of jurors chosen by lot from a pool of citizens over 30 years of age; the pool itself was chosen annually and also by lot.

In 411 BC, exploiting the unrest created in Athens by the Peloponnesian War with Sparta, a group known as the Four Hundred seized control of Athens and established an oligarchy. Less than a year later, the Four Hundred were overthrown and democracy was restored. Nine decades later, in 321, Athens was subjugated by its northern neighbour Macedonia, which introduced property qualifications that effectively excluded many ordinary Athenians from the *dēmos*. In 146 BC what remained of Athenian democracy was extinguished by the conquering Romans.

THE ROMAN REPUBLIC

At about the same time that popular government was introduced in Greece, it also appeared on the Italian Peninsula in the city of Rome. The Romans called their system a *rēspublica,* or "republic," from the Latin *res,* meaning "thing" or "affair," and *publicus* or *publica*, meaning "public"—thus, a republic was the thing that belonged to the Roman people, the *populus romanus*.

Like Athens, Rome was originally a city-state. Although it expanded rapidly by conquest and annexation far beyond its original borders to encompass the entire Mediterranean world and much of western Europe, its government remained essentially that of a moderately large city-state. Indeed, throughout the republican era (until roughly the end of the 1st century BC), Roman assemblies were held in the very small Forum at the centre of the city.

Although Roman citizenship was conferred by birth, it was also granted by naturalization and by manumission of slaves. As the Roman Republic expanded, it conferred citizenship in varying degrees to many of those within its enlarged boundaries. Because Roman assemblies continued to meet in the Forum, most citizens who did not live in or near the city itself were effectively excluded from the *dēmos.*

The Romans created a political structure so complex that later democratic leaders chose not to emulate it. They used not only an extremely powerful Senate but also four assemblies, each called *comitia* ("assembly") or *concilium* ("council"). The Comitia Curiata was composed of 30 curiae, or local groups, drawn from three ancient *tribus*, or tribes; the Comitia Centuriata consisted of 193 centuries, or military units; the Concilium Plebis was drawn from the ranks of the plebes, or plebeians (common people); and the Comitia Tributa, like the Athenian Assembly, was open to all citizens. In all the assemblies, votes were counted by units (centuries or tribes) rather than by individuals.

Although they collectively represented all Roman citizens, the assemblies were not sovereign. Throughout the entire period of the republic, the Senate—an institution inherited from the earlier era of the Roman monarchy—continued to exercise great power. Senators were chosen indirectly by the Comitia Centuriata; during the monarchy, they were drawn exclusively from the privileged patrician class, though later, during the republic, members of certain plebeian families were also admitted.

THE ITALIAN REPUBLICS FROM THE 12TH CENTURY TO THE RENAISSANCE

"Constitutional oligarchies." After the western Roman Empire collapsed in 476, the Italian Peninsula broke up into a congeries of smaller political entities, some of which, about six centuries later, developed into more or less independent city-states. These city-states inaugurated systems of government based on wider—though not fully popular—participation and on the election of leaders for limited periods of time. Such governments flourished for two centuries or more in a number of cities, including Venice, Florence, Siena, and Pisa.

Drawing on Latin rather than Greek, the Italians called their city-states republics, not democracies. Membership in the *dēmos* was at first restricted mainly to the nobility and large landowners. In some republics in the first half of the 13th century, however, some groups from the lower social and economic classes—such as the newly rich, small merchants and bankers, and skilled craftsmen—began to demand the right to participate in government at some level. Because they were more numerous than the upper classes and because they threatened (and sometimes carried out) violent uprisings, some of these groups were successful. Even with these additions, however, the *dēmos* in the republics remained only a tiny fraction of the total population, ranging from 12 percent in 14th-century Bologna to 2 percent or less in 15th- and 16th-century Venice. Thus, whether judged by the standards of Classical Greece or those of Europe and the United States in the 18th century and later, the Italian republics were not democracies. A more accurate characterization, proposed by the historian Lauro Martines, is "constitutional oligarchies."

A democratic dilemma. The Greeks, the Romans, and the leaders of the Italian republics were pioneers in creating popular governments, and their philosophers and commentators exercised enormous influence on later political thought. Yet their political institutions were not emulated by the later founders of democratic governments in the nation-states of northern Europe and North America. As the expansion of Rome had already demonstrated, these institutions were simply not suited to political associations significantly larger than the city-state.

The enormous difference in size between a city-state and a nation-state indicates a fundamental dilemma. By limiting the size of a political association to that of a city-state, citizens can in principle ensure their capacity to influence directly the conduct of their government—*e.g.*, by participating in an assembly. But limiting size comes at a cost: important problems—notably defense and the regulation of trade and finance—will remain beyond the capacity of the government to deal with effectively. Alternatively, by increasing the size of the association—*i.e.*, by enlarging its geographic area and population—citizens can increase the capacity of the government to deal with important problems, but only at the cost of reducing their opportunities to influence the government directly through assemblies or other means.

Many city-states responded to this dilemma by forming alliances or confederations with other city-states and with larger political associations. But the problem would not finally be solved until the development of representative government, which first appeared in northern Europe in the 18th century.

TOWARD REPRESENTATIVE DEMOCRACY: EUROPE AND NORTH AMERICA TO THE 19TH CENTURY

Regional developments. *Continental Europe.* About AD 800, freemen and nobles in various parts of northern Continental Europe began to participate directly in local assemblies, to which were later added regional and national assemblies consisting of representatives, some or all

of whom came to be elected. In the mountain valleys of the Alps, such assemblies developed into self-governing cantons, leading eventually to the founding of the Swiss Confederation in the 13th century. By 900, local assemblies of Vikings were meeting in many areas of Scandinavia. Eventually the Vikings realized that to deal with certain larger problems they needed more-inclusive associations, and in Norway, Sweden, and Denmark regional assemblies developed. In 930 Viking descendants in Iceland created the first example of what today would be called a national assembly or parliament—the Althing.

The English Parliament

England. Among the assemblies created in Europe during the Middle Ages, the one that most profoundly influenced the development of representative government was the English Parliament. Parliament grew out of councils that were called by kings for the purpose of redressing grievances and for exercising judicial functions. In time, it began to deal with important matters of state, notably the raising of revenues needed to support the policies of the monarch. As its judicial functions were increasingly delegated to courts, it gradually evolved into a legislative body. By the end of the 15th century, the English system displayed some of the basic features of modern parliamentary government: for example, the enactment of laws now required the passage of bills by both houses of Parliament and the formal approval of the monarch.

By about 1800, significant powers, notably including powers related to the appointment and tenure of the prime minister, had shifted to Parliament. This development was strongly influenced by the emergence of political factions during the early years of the 18th century. These factions, known as Whigs and Tories, later became full-fledged parties. To king and Parliament alike it became increasingly apparent that laws could not be passed nor taxes raised without the support of a Whig or Tory leader who could muster a majority of votes in the House of Commons. To gain that support, the monarch was forced to select as prime minister the leader of the majority party in the Commons and to accept the leader's suggestions for the composition of the cabinet. By 1830 the constitutional principle that the choice of prime minister, and thus the cabinet, reposed with the House of Commons had become firmly entrenched in the (unwritten) British Constitution.

Parliamentary government in Britain was not yet a democratic system, however. Mainly because of property requirements, the franchise was held by only about 5 percent of the British population over 20 years of age. The Reform Act of 1832, which is generally viewed as a historic threshold in the development of parliamentary democracy in Britain, extended the suffrage to about 7 percent of the adult population. It would require further acts of Parliament in 1867, 1884, and 1918 to achieve universal male suffrage and one more law, enacted in 1928, to secure the right to vote for all adult women.

The United States. Whereas the feasibility of representative government was demonstrated by the development of Parliament, the possibility of joining representation with democracy first became fully evident in the governments of the British colonies of North America and later in the founding of the United States of America.

Conditions in colonial America favoured the limited development of a system of representation more broadly based than the one in use in Britain. These conditions included the vast distance from London, which forced the British government to grant significant autonomy to the colonies; the existence of colonial legislatures in which representatives in at least one house were elected by voters; the expansion of the suffrage, which in some colonies came to include most adult white males; the spread of property ownership, particularly in land; and the strengthening of beliefs in fundamental rights and popular sovereignty, including the belief that the colonists should not have to pay taxes to a government in which they were not represented ("no taxation without representation").

Because of the new country's large population and enormous size, democratic government was possible at the federal, state, and territorial levels only through representatives. In smaller associations, however, a direct assembly of citizens was both feasible and desirable. In many New England towns, for example, citizens assembled in meetings, Athenian style, to discuss and vote on local matters.

Thus, the United States provided new answers to question 1—What is the appropriate unit or association within which a democratic government should be established?—and question 3—How are citizens to govern? Yet, the American answer to question 2—Who should constitute the *dēmos*?—though radical in its time, was by later standards highly unsatisfactory. Even as the suffrage was broadly extended among adult white males, it continued to exclude large segments of the adult population, such as women, slaves, many freed blacks, and Native Americans.

Democracy or republic? Is *democracy* the most appropriate name for a large-scale representative system such as that of the early United States? At the end of the 18th century, the history of the terms whose literal meaning is "rule by the people"—*democracy* and *republic*—left the answer unclear. Both terms had been applied to the assembly-based systems of Greece and Rome, though neither assigned legislative powers to representatives elected by members of the *dēmos*. When the United States Constitutional Convention was held in 1787, terminology was still unsettled. Not only were *democracy* and *republic* used more or less interchangeably in the colonies, but no established term existed for a representative government "by the people."

Given the existing confusion over terminology, it is not surprising that the framers employed various terms to describe the novel government they proposed. In "Federalist 10," one of 85 essays by James Madison, Alexander Hamilton, and John Jay known collectively as the Federalist papers, Madison defined a "pure democracy" as "a society consisting of a small number of citizens, who assemble and administer the government in person," and a republic as "a government in which the scheme of representation takes place." Thus, for Madison *democracy* meant direct democracy, and *republic* meant representative government.

"Federalist 10"

In November 1787, James Wilson, one of the signers of the Declaration of Independence, proposed a new classification. He wrote,

> [T]he three species of governments are the monarchical, aristocratical and democratical. In a monarchy, the supreme power is vested in a single person: in an aristocracy . . . by a body not formed upon the principle of representation, but enjoying their station by descent, or election among themselves, or in right of some personal or territorial qualifications; and lastly, in a democracy, it is inherent in a people, and is exercised by themselves or their representatives.

At the Virginia ratifying convention some months later, John Marshall, the future chief justice of the Supreme Court, declared that the "Constitution provided for 'a well regulated democracy' where no king, or president, could undermine representative government." The political party that he organized in cooperation with Thomas Jefferson, the future third president of the United States, was named the Democratic-Republican Party; the party adopted its present name, the Democratic Party, in 1844.

Following his visit to the United States in 1831–32, the French political scientist Alexis de Tocqueville asserted in no uncertain terms that the country he had observed was a democracy—indeed, the world's first representative democracy, where the fundamental principle of government was "the sovereignty of the people." Tocqueville's estimation of the American system reached a wide audience in Europe and beyond through his monumental four-volume study *Democracy in America* (1835–40).

Solving the dilemma. Thus, by the end of the 18th century both the idea and the practice of democracy had been profoundly transformed. Political theorists and statesmen now recognized that the nondemocratic practice of representation could be used to make democracy practicable in the large nation-states of the modern era. Representation, in other words, was the solution to the ancient dilemma between enhancing the ability of political associations to deal with large-scale problems and preserving the opportunity of citizens to participate in government.

To some of those steeped in the older tradition, the union of representation and democracy seemed a marvelous and

epochal invention. In the early 19th century the French author Destutt de Tracy, the inventor of the term *idéologie* ("ideology"), insisted that representation had rendered obsolete the doctrines of both Montesquieu and Jean-Jacques Rousseau, both of whom had denied that representative governments could be genuinely democratic. In 1820 the English philosopher James Mill proclaimed "the system of representation" to be "the grand discovery of modern times" in which "the solution of all the difficulties, both speculative and practical, will perhaps be found." One generation later Mill's son, the philosopher John Stuart Mill, concluded in his *Considerations on Representative Government* (1861) that "the ideal type of a perfect government" would be both democratic and representative.

New answers to old questions. *Suffrage.* In the 19th and 20th centuries there were important changes in the prevailing answers to some of the other fundamental questions mentioned earlier. One important development concerned question 2—Who should constitute the *dēmos?* In the 19th century, property requirements for voting were reduced and finally removed. The exclusion of women from the *dēmos* was increasingly challenged—not least by women themselves. Beginning with New Zealand in 1893, more and more countries granted women the suffrage and other political rights, and by the mid-20th century women were full and equal members of the *dēmos* in almost all countries that considered themselves democratic.

Although the United States granted women the right to vote in 1920, another important exclusion continued for almost half a century: African Americans were prevented, by both legal and illegal means, from voting and other forms of political activity, primarily in the South but also in other areas of the country. Not until after the passage and enforcement of the Civil Rights Act of 1964 were they effectively admitted into the American *dēmos.* Native Americans suffered similar forms of discrimination.

The Civil Rights Act

Thus, in the 19th and 20th centuries the *dēmos* was gradually expanded to include all adult citizens. By the mid-20th century, no system whose *dēmos* did not include all adult citizens could properly be called democratic.

Factions and parties. In many of the city-state democracies and republics, part of the answer to question 3—What political institutions are necessary for governing?—consisted of factions, including both informal groups and organized political parties. Much later, representative democracies in several countries developed political parties for selecting candidates for parliament and for organizing parliamentary support for (or opposition to) the prime minister and his cabinet. Nevertheless, at the end of the 18th century leading political theorists continued to regard factions as a profound danger to democracies and republics. This view was also common at the United States Constitutional Convention.

Factions are dangerous, it was argued, for at least two reasons. First, a faction is by definition a group whose interests are in conflict with the general good. Second, historical experience shows that, prior to the 18th century, the existence of factions in a democracy or republic tended to undermine the stability of its government. The "instability, injustice, and confusion introduced into the public councils" by factionalism, Madison wrote, have been "the mortal diseases under which popular governments have everywhere perished."

Interestingly, Madison used the presumed danger of factions as an argument in favour of adopting the new constitution. Because the United States, in comparison with previous republics, would have many more citizens and vastly more territory, the diversity of interests among its population would be much greater, making the formation of large or powerful factions less likely. Similarly, the exercise of government power by representatives rather than directly by the people would "refine and enlarge the public views, by passing them through the medium of a chosen body of citizens, whose wisdom may best discern the true interest of their country."

Madison eventually abandoned his belief in the essential perniciousness of factions. Political parties, he came to believe, were not only legally possible, necessary, and inevitable—they were also desirable. They were legally possible because of the rights and liberties provided for in the constitution. They were necessary in order to defeat the Federalists, whose centralizing policies Madison, Jefferson, and many others strongly opposed. Because parties were both possible and necessary, they would inevitably be created. Finally, parties were also desirable, because, by helping to mobilize voters throughout the country and in the legislative body, they enabled the majority to prevail over the opposition of a minority.

This view came to be shared by political thinkers in other countries in which democratic forms of government were developing. By the end of the 19th century, it was nearly universally accepted that the existence of independent and competitive political parties is an elementary standard that every democracy must meet.

Majority rule, minority rights, majority tyranny. The fear of "majority tyranny" was a common theme in the 17th century and later, even among those who were sympathetic to democracy. Given the opportunity, it was argued, a majority would surely trample on the fundamental rights of minorities. Property rights were perceived as particularly vulnerable, since presumably any majority of citizens with little or no property would be tempted to infringe the rights of the propertied minority. Such concerns were shared by Madison and other delegates at the Convention—including Benjamin Franklin, who once observed that "democracy is two wolves and a lamb voting on what to have for lunch."

The fear of majority tyranny was eased and finally abandoned after democratic leaders, including Madison himself, realized that unrestrained majority rule could be blocked by numerous legal barriers, none of which would be clearly inconsistent with basic democratic principles. Thus, a bill of rights could be incorporated into the constitution; a supermajority of votes could be required for constitutional amendments and other important kinds of legislation; the executive, legislative, and judicial powers of government could be divided into separate branches; and an independent judiciary could be given the power to declare laws or policies unconstitutional.

Although political theorists continue to disagree about the best means to effect majority rule in democratic systems, it is evident that majorities cannot legitimately abridge the fundamental rights of citizens. In short, because democracy is not only a political system of "rule by the people" but necessarily also a system of rights, a government that infringes these rights is to that extent undemocratic.

Democracy as a system of rights

THE SPREAD OF DEMOCRACY IN THE 20TH CENTURY

During the 20th century, the number of countries possessing the basic political institutions of representative democracy increased significantly. At the beginning of the 21st century, more than one-third of the world's nominally independent countries possessed democratic institutions comparable to those of the English-speaking countries and the older democracies of Continental Europe. In an additional one-sixth of the world's countries, these institutions, though somewhat defective, nevertheless provided historically high levels of democratic government. What accounted for this rapid expansion of democracy?

Failures of nondemocratic systems. A significant part of the explanation is that all the main alternatives to democracy—whether of ancient or of modern origins—suffered political, economic, diplomatic, and military failures that greatly lessened their appeal. With the victory of the Allies in World War I, the ancient systems of monarchy, aristocracy, and oligarchy ceased to be legitimate. Following the military defeat of Italy and Germany in World War II, the newer alternative of fascism was likewise discredited, as was Soviet-style communism after the economic and political collapse of the Soviet Union in 1990–91. Similar failures contributed to the gradual disappearance of military dictatorships in Latin America in the 1980s and '90s.

Market economies. Accompanying these ideological and institutional changes were changes in economic institutions. Highly centralized economies under state control had enabled political leaders to use their access to economic resources to reward their allies and punish their critics. As these systems were displaced by more decentralized

market economies, the power and influence of top government officials declined. In addition, some of the conditions that were essential to the successful functioning of market economies also contributed to the development of democracy: access to reliable information, relatively high levels of education, ease of personal movement, and the rule of law. As market economies expanded and as middle classes grew larger and more influential, popular support for such conditions increased, often accompanied by demands for further democratization.

Economic well-being. As the economic well-being of large segments of the world's population gradually improved, so too did the likelihood that newly established democratic institutions would survive and flourish. In general, citizens in democratic countries with persistent poverty are more susceptible to the appeals of antidemocratic demagogues who promise simple and immediate solutions to their country's economic problems. Accordingly, widespread economic prosperity in a country greatly increases the chances that democratic government will succeed.

Correlation of democracy and prosperity

Political culture. During the 20th century, democracy continued to exist in some countries despite periods of acute diplomatic, military, economic, or political crisis, such as occurred during the early years of the Great Depression. The survival of democratic institutions in these countries is attributable in part to the existence in their societies of a culture of widely shared democratic beliefs and values. In countries where democratic culture is weak or absent, such as the Weimar Republic of Germany in the years following World War I, democracy is much more vulnerable, and periods of crisis are more likely to lead to a reversion to a nondemocratic regime.

CONTEMPORARY DEMOCRATIC SYSTEMS

Differences among democratic countries in historical experience, size, ethnic and religious composition, and other factors have resulted in significant differences in their political institutions. Political scientists have used these differences to identify a few basic kinds of democratic political system.

Presidential and parliamentary systems. Whereas versions of the U.S. presidential system were frequently adopted in Latin America, Africa, and elsewhere in the developing world, as European countries democratized, they adopted versions of the English parliamentary system, which made use of both a prime minister responsible to parliament and a ceremonial head of state, who may be a hereditary monarch—as in the Scandinavian countries, The Netherlands, and Spain—or a president chosen by parliament—as in Israel. A notable exception is France, which in its fifth constitution, adopted in 1958, combined its parliamentary system with a presidential one.

Unitary and federal systems. In most older European and English-speaking democracies, political authority inheres in the central government, which is constitutionally authorized to determine the limited powers, as well as the geographic boundaries, of subnational associations such as states and regions. Such unitary systems contrast markedly with federal systems, in which authority is constitutionally divided between the central government and the governments of relatively autonomous subnational entities. Democratic countries that have adopted federal systems include—in addition to the United States—Switzerland, Germany, Austria, Spain, Canada, and Australia. The world's most populous democratic country, India, also has a federal system.

Proportional and winner-take-all systems. Electoral arrangements vary enormously. Some democratic countries divide their territories into electoral districts, each of which is entitled to a single seat in the legislature, the seat being won by the candidate who gains the most votes—hence the terms *first past the post* in Britain and *winner take all* in the United States. As critics of this system point out, in districts contested by more than two candidates, it is possible to gain the seat with less than a strict majority of votes (50 percent plus one). As a result, a party that receives only a minority of votes in the entire country could win a majority of seats in the legislature. Systems of proportional representation are designed to ensure a closer correspondence between the proportion of votes cast for a party and the proportion of seats it receives. With few exceptions, Continental European countries have adopted some form of proportional representation, as have Ireland, Australia, New Zealand, and Japan. Winner-take-all systems remain in the United States, Canada, and, for parliamentary elections, Britain.

Two-party and multiparty systems. Because proportional representation does not favour large parties over smaller ones, as does the winner-take-all system, in countries with proportional representation there are almost always three or more parties represented in the legislature, and a coalition government consisting of two or more parties is ordinarily necessary to win legislative support for the government's policies. Thus the prevalence of proportional representation effectively ensures that coalition governments are the rule in democratic countries; governments consisting of only two parties, such as that of the United States, are extremely rare.

Majoritarian and consensual systems. Because of differences in electoral systems and other factors, democratic countries differ with respect to whether laws and policies can be enacted by a single, relatively cohesive party with a legislative majority, as is ordinarily the case in Britain and Japan, or instead require consensus among several parties with diverse views, as in Switzerland, The Netherlands, Sweden, Italy, and elsewhere. Political scientists and others disagree about which of the two types of system, majoritarian or consensual, is more desirable. Critics of consensual systems argue that they allow a minority of citizens to veto policies they dislike and that they make the tasks of forming governments and passing legislation excessively difficult. Supporters contend that consensual arrangements produce comparatively wider public support for government policies and even help to increase the legitimacy and perceived value of democracy itself.

The theory of democracy

DEMOCRATIC IDEAS FROM PERICLES TO RAWLS

The "theory of democracy" refers to the study of the nature and ultimate value of democracy, an endeavour that is both empirical and philosophical. Not surprisingly, the first important contributions to the theory of democracy were made by the ancient Greeks, and particularly by political leaders and philosophers of Athens.

Pericles. In a funeral oration in 430 BC for those who had fallen in the Peloponnesian War, the Athenian leader Pericles described democratic Athens as "the school of Hellas." Among the city's many exemplary qualities, he declared, was its constitution, which "favors the many instead of the few; this is why it is called a democracy." Pericles continued: "If we look to the laws, they afford equal justice to all in their private differences; if to social standing, advancement in public life falls to reputation for capacity, class considerations not being allowed to interfere with merit; nor again does poverty bar the way; if a man is able to serve the state, he is not hindered by obscurity of his condition. The freedom which we enjoy in our government extends also to our ordinary life."

"The school of Hellas"

Aristotle. A century later, Aristotle discussed democracy in terms that would become highly influential in comparative studies of political systems. At the heart of his approach is the notion of a "constitution." Aristotle identifies three kinds of ideal constitution—each of which describes a situation in which those who rule pursue the common good—and three corresponding kinds of perverted constitution—each of which describes a situation in which those who rule pursue narrow and selfish goals. The three kinds of constitution, both ideal and perverted, are differentiated by the number of persons they allow to rule. Thus "rule by one" is monarchy in its ideal form and tyranny in its perverted form; "rule by the few" is aristocracy in its ideal form and oligarchy in its perverted form; and "rule by the many" is "polity" in its ideal form and democracy in its perverted form.

Aristotle's general scheme prevailed for more than two millennia, though his unsympathetic and puzzling definition of democracy did not.

Locke. Nearly 20 centuries after Aristotle, the English philosopher John Locke adopted the essential elements of the Aristotelian classification of constitutions in his *Second Treatise of Civil Government* (1690). Unlike Aristotle, however, Locke was an unequivocal supporter of political equality, individual liberty, democracy, and majority rule. Although his work was naturally rather abstract and not particularly programmatic, it provided a powerful philosophical foundation for much later democratic theorizing and political programs.

The legitimacy of government. According to Locke, in the hypothetical "state of nature" that precedes the creation of human societies, men live "equal one amongst another without subordination or subjection," and they are perfectly free to act and to dispose of their possessions as they see fit, within the bounds of natural law. From these and other premises Locke draws the conclusion that political society—*i.e.*, government—insofar as it is legitimate, represents a social contract among those who have "consented to make one Community or Government . . . wherein the Majority have a right to act and conclude the rest." These two ideas—the consent of the governed and majority rule—became central to all subsequent theories of democracy. For Locke they are inextricably connected: no government is legitimate unless it enjoys the consent of the governed, and that consent cannot be rendered except through majority rule.

Locke differentiates the various forms of government on the basis of where the people choose to place the power to make laws. If the people retain the legislative power for themselves, together with the power to appoint those who execute the laws, then "the Form of the Government is a perfect Democracy." If they put the power "into the hands of a few select Men, and their Heirs or Successors, . . . then it is an Oligarchy: Or else into the hands of one Man, and then it is a Monarchy." Although it relies on the traditional categories, Locke's analysis is far more subversive of nondemocratic forms of government than it appears to be. For whatever the form of government, the ultimate source of sovereign power is the people, and all legitimate government must rest on their consent. Therefore, if a government abuses its trust and violates the people's fundamental rights—particularly the right to property—the people are entitled to rebel and replace that government with another to whose laws they can willingly give their consent. Moreover, it is the people themselves who are entitled to judge whether the government has abused its trust. Although he does not use the term, Locke thus unambiguously affirms the right of revolution against a despotic government.

The Declaration of Independence

Less than a century later, Locke's views were echoed in the famous words of the United States Declaration of Independence:

> We hold these truths to be self-evident, that all men are created equal, that they are endowed by their Creator with certain unalienable Rights, that among these are Life, Liberty, and the pursuit of Happiness. That to secure these rights, Governments are instituted among Men, deriving their just powers from the consent of the governed, that whenever any Form of Government becomes destructive of these ends, it is the Right of the People to alter or abolish it . . . .

Answers to fundamental questions. Although Locke's ideas were radical in his time, his answers to questions 1 through 3 would need further elaboration, and even some alteration, as the theory and practice of democracy continued to develop.

Regarding question 1—What is the appropriate association within which a democratic government should be established?—Locke clearly intended his conclusions to apply to England as a whole and presumably also to other nation-states. Departing from views that still prevailed among political philosophers of his time, Locke held that democracy did not require a small political unit, such as a city-state, in which all members of the *dēmos* could participate in government directly.

Regarding question 2—Who should constitute the *dēmos*?—Locke believed, along with almost everyone else who had expressed an opinion on the issue, that children should not enjoy the full rights of citizenship, though he maintained that parents are morally obliged to respect their children's rights as human beings. With almost no substantive argument, Locke also adopted the traditional view that women should be excluded from the *dēmos,* though he insisted that they retain all other fundamental rights.

Unlike the men of Athens or the small male aristocracy of Venice, obviously the men of England could not govern themselves directly in an assembly. In this case, then, the answer to question 3—What political institutions are necessary for governing?—would have to include the use of representatives chosen by the people. Yet Locke provided little guidance as to the form a representative government might take.

Montesquieu. The French political theorist Montesquieu, through his masterpiece *The Spirit of the Laws* (1748), strongly influenced his younger contemporary Rousseau and many of the American Founding Fathers, including John Adams, Jefferson, and Madison. Rejecting Aristotle's classification, Montesquieu distinguishes three ideal types of government: monarchy, "in which a single person governs by fixed and established laws"; despotism, "in which a single person directs everything by his own will and caprice"; and republican (or popular) government, which may be of two types, depending on whether "the body, or only a part of the people, is possessed of the supreme power," the former being a democracy, the latter an aristocracy.

According to Montesquieu, a necessary condition for the existence of a republican government, whether democratic or aristocratic, is that the people in whom supreme power is lodged possess the quality of "public virtue," meaning that they are motivated by a desire to achieve the public good. Although public virtue may not be necessary in a monarchy and is certainly absent in despotic regimes, it must be present to some degree in aristocratic republics and to a large degree in democratic republics. Sounding a theme that would be loudly echoed in Madison's "Federalist 10," Montesquieu asserts that without strong public virtue, a democratic republic is likely to be destroyed by conflict between various "factions," each pursuing its own narrow interests at the expense of the broader public good.

Hume. The destructive power of factions was also emphasized by the Scottish philosopher David Hume, whose influence on Madison was perhaps even greater than Montesquieu's. For it was from Hume that Madison seems to have acquired a view about factions that turned the issue of the desirability of larger political associations—*i.e.*, those larger than the city-state—on its head. For the purpose of diminishing the destructive potential of factionalism, so Hume and Madison argued, bigger is in fact better, because in bigger associations each representative must look after a greater diversity of interests.

Rousseau. When compared with Locke, Jean-Jacques Rousseau sometimes seems the more radical democrat, though a close reading of his work shows that, in important respects, Rousseau's conception of democracy is narrower than Locke's. Indeed, in his most influential work of political philosophy, *The Social Contract* (1762), Rousseau asserts that democracy is incompatible with representative institutions, a position that renders it all but irrelevant to nation-states. The sovereignty of the people, he argues, can be neither alienated nor represented. But if representation is incompatible with democracy, and if direct democracy is the only legitimate form of government, then no nation-state of Rousseau's time or any other can have a legitimate government. Furthermore, according to Rousseau, if a political association that is small enough to practice direct democracy, such as a city-state, were to come into existence, it would inevitably be subjugated by larger nation-states and thereby cease to be democratic.

The Social Contract

For these and other reasons, Rousseau was pessimistic about the prospects of democracy. Adopting a view common among critics of democracy in his time, Rousseau also held that "there is no government so subject to civil wars and intestine agitations as democratic or popular government." In a much-cited passage, he declares that, "were there a people of gods, their government would be democratic. So perfect a government is not for men."

Mill. In his work *On Liberty* (1859) John Stuart Mill ar-

gued on utilitarian grounds that individual liberty cannot be legitimately infringed—whether by government, society, or individuals—except in cases where the individual's action would cause harm to others. This principle provided a philosophical foundation for some of the basic freedoms essential to a democracy—such as freedom of association—and undermined the legitimacy of paternalistic laws—such as those requiring temperance—which in Mill's view treated adult citizens like children. In the area of what he called the liberty of thought and discussion, another freedom crucial to democracy, Mill argued, also on utilitarian grounds, that legal restrictions on the expression of opinion are never justified. The "collision of adverse opinions," he contended, is a necessary part of any society's search for the truth. In another work, *Considerations on Representative Government* (1861), Mill set forth in a lucid and penetrating manner many of the essential features of the new type of government, which had not yet emerged in Continental Europe and was still incomplete in important respects in the United States. In this work he also advanced a powerful argument on behalf of woman suffrage—a position that virtually all previous political philosophers (all of whom were male) had ignored or rejected.

Dewey. According to the American philosopher John Dewey, democracy is the most desirable form of government because it alone provides the kinds of freedom necessary for individual self-development and growth—the freedom to exchange ideas and opinions, the freedom to form associations to pursue common goals, and the freedom to determine and pursue one's own conception of the "good life." Democracy is more than a form of government, however; as Dewey remarks in *Democracy and Education* (1916), it is also a "mode of associated life" in which citizens cooperate with each other to solve their common problems through rational means in a spirit of mutual respect and good will. Moreover, the political institutions of any democracy, according to Dewey, should not be viewed as the perfect and unchangeable creations of visionary statesmen of the past; rather, they should be constantly subject to criticism and improvement as historical circumstances and the public interest change.

Democracy and Education

Participation in a democracy as Dewey conceived it requires critical and inquisitive habits of mind, an inclination toward cooperation, and a feeling of public spiritedness and a desire to achieve the common good. Because these habits and inclinations must be inculcated from a young age, Dewey placed great emphasis on education. His contributions to both the theory and practice of education were enormously influential in the United States in the 20th century.

Dewey offered few concrete proposals regarding the form that democratic institutions should take. Nevertheless, in *The Public and Its Problems* (1927) and other works, he contended that individuals cannot develop to their fullest potential except in a social democracy, or a democratic welfare-state. Accordingly, he held that democracies should possess strong regulatory powers. He also insisted that among the most important features of a social democracy should be the right of workers to participate directly in the management of the firms in which they are employed.

Rawls. From the time of Mill until about the mid-20th century, most philosophers who defended democratic principles did so largely on utilitarian grounds—*i.e.*, they argued that systems of government that are democratic in character are more likely than other systems to produce a greater amount of happiness (or well-being) for a greater number of people. Such justifications, however, were traditionally vulnerable to the objection that they could be used to support forms of government in which the greater happiness of the majority is achieved by unfairly neglecting the rights and interests of a minority.

In *A Theory of Justice* (1971), the American philosopher John Rawls attempted to develop a nonutilitarian justification of a democratic political order characterized by fairness, equality, and individual rights. Reviving the notion of a social contract, which had been dormant since the 18th century, he imagined a hypothetical situation in which a group of rational individuals are rendered ignorant of all social and economic facts about themselves—including facts about their race, sex, religion, education, intelligence, talents or skills, and even their conception of the good life—and then asked to decide what general principles should govern the political institutions under which they live. From behind this "veil of ignorance," Rawls argues, such a group would unanimously reject utilitarian principles—such as "political institutions should aim to maximize the happiness of the greatest number"—because no member of the group could know whether he belonged to a minority whose rights and interests might be neglected under institutions justified on utilitarian grounds.

The "veil of ignorance"

Instead, reason and self-interest would lead the group to adopt principles such as the following: (1) Everyone should have a maximum and equal degree of liberty, including all the liberties traditionally associated with democracy. (2) Everyone should have an equal opportunity to seek offices and positions that offer greater rewards of wealth, power, status, or other social goods. (3) The distribution of wealth in society should be such that those who are least well-off are better off than they would be under any other distribution, whether equal or unequal. (Rawls holds that some inequality in the distribution of wealth may be necessary to achieve higher levels of productivity. It is therefore possible to imagine unequal distributions of wealth in which those who are least well-off are better off than they would be under an equal distribution.) These three principles amount to an egalitarian form of democratic liberalism. Rawls is accordingly regarded as the leading philosophical defender of the modern democratic capitalist welfare state.

THE VALUE OF DEMOCRACY

Why should "the people" rule? Is democracy really superior to any other form of government? Although a full exploration of this issue is beyond the scope of this article, history—particularly 20th-century history—demonstrates that democracy uniquely possesses a number of features that most people, whatever their basic political beliefs, would consider desirable: (1) Democracy helps to prevent rule by cruel and vicious autocrats. (2) Modern representative democracies do not fight wars with one another. (3) Countries with democratic governments tend to be more prosperous than countries with nondemocratic governments. (4) Democracy tends to foster human development—as measured by health, education, personal income, and other indicators—more fully than other forms of government do.

Other features of democracy also would be considered desirable by most people, though some would regard them as less important than features 1 through 4 above: (5) Democracy helps people to protect their fundamental interests. (6) Democracy guarantees its citizens fundamental rights that nondemocratic systems do not, and cannot, grant. (7) Democracy ensures its citizens a broader range of personal freedoms than other forms of government do.

Finally, there are some features of democracy that some people—the critics of democracy—would not consider desirable at all, though most people, upon reflection, would regard them as at least worthwhile: (8) Only democracy provides people with a maximum opportunity to live under laws of their own choosing. (9) Only democracy provides people with a maximum opportunity to take moral responsibility for their choices and decisions about government policies. (10) Only in a democracy can there be a relatively high level of political equality.

These advantages notwithstanding, there have been critics of democracy since ancient times. Perhaps the most enduring of their charges is that most people are incapable of participating in government in a meaningful or competent way because they lack the necessary knowledge, intelligence, wisdom, experience, or character. According to Plato, for example, the best government would be an aristocracy of philosopher-kings whose rigorous intellectual and moral training would make them uniquely qualified to rule. The view that the people as a whole are incapable of governing themselves has been espoused not only by kings and aristocratic rulers but also by political theorists (Plato

Criticisms of democracy

foremost among them), religious leaders, and other authorities. The view was prevalent in one form or another throughout the world during most of recorded history until the early 20th century.

No doubt there will be critics of democracy for as long as democratic governments exist. The extent of their success in winning adherents and promoting the creation of nondemocratic regimes will depend on how well democratic governments meet the new challenges and crises that are all but certain to occur.

Problems and challenges

At the beginning of the 21st century, democracy faced a number of challenges, some of which had been problems of long standing, others of which were more recent.

INEQUALITY OF RESOURCES

Although market economies encouraged the spread of democracy, in countries where they were not sufficiently regulated such economies eventually produced large inequalities in economic and social resources, from wealth and income to education and social status. Because those with greater resources naturally tended to use them to influence the political system to their advantage, the existence of such inequalities constituted a persistent obstacle to the achievement of a satisfactory level of political equality.

IMMIGRATION

After World War II, immigration to the countries of western Europe, Australia, and the United States, both legal and illegal, increased dramatically. Differences in language, culture, and appearance between immigrant groups and the citizens of the host country, as well as the usually widespread perception that immigrants take jobs away from citizens and use expensive social services, made immigration a hotly debated issue in many countries. In some instances, anti-immigrant sentiment contributed to the rise of radical political parties and movements, some of which promoted racist or neofascist doctrines that were hostile not only to immigrants but also to fundamental political and human rights and even to democracy itself.

TERRORISM

Acts of terrorism committed within democratic countries or against their interests in other parts of the world occurred with increasing frequency beginning in the 1970s. In response to such events, democratic governments adopted various measures designed to enhance the ability of police and other law-enforcement agencies to protect their countries against terrorism. Some of these initiatives entailed new restrictions on citizens' civil and political liberties and were accordingly criticized as unconstitutional or otherwise inconsistent with democratic principles. In the early 21st century it remained to be seen whether democratic governments could strike a satisfactory balance between the sometimes conflicting imperatives of ensuring security and preserving democracy.

INTERNATIONAL SYSTEMS

At the end of the 18th century, in response to the dilemma of size described earlier, the focus of both the theory and the practice of democracy shifted from the small association of the city-state to the far larger nation-state. Although their increased size enabled democracies to solve more of the problems they confronted, there remained some problems that not even the largest democracy could solve by itself. To address these problems several international organizations were established after World War II, most notably the United Nations (1945), and their numbers and responsibilities grew rapidly through the rest of the 20th century.

The United Nations

These organizations posed two related challenges to democracy. First, by shifting ultimate control of a country's policies in a certain area to the international level, they reduced to a corresponding extent the influence that citizens could exert on such policies through democratic means. Second, all international organizations, even those that were formally accountable to national governments, lacked the political institutions of representative democracy. How could these institutions be made democratic—or at least more democratic?

TRANSITION, CONSOLIDATION, BREAKDOWN

The problems and challenges facing democracy were particularly acute in countries that became democratic in the late 20th and early 21st centuries. Obstacles in the path of a successful consolidation of democratic institutions include economic problems such as widespread poverty, unemployment, massive inequalities in income and wealth, rapid inflation, and low or negative rates of economic growth. Countries at low levels of economic development also usually lack a large middle class and a well-educated population. In many of these countries, the division of the population into antagonistic ethnic, racial, religious, or linguistic groups has made it difficult to manage political differences peacefully. In others, extensive government intervention in the economy, along with other factors, has resulted in the widespread corruption of government officials. Many countries also lack an effective legal system, making civil rights highly insecure and allowing for abuse by political elites and criminal elements. In these countries the idea of the rule of law is not well established in the prevailing political culture, and in other respects the political culture of these countries does not inculcate in citizens the kinds of beliefs and values that could support democratic institutions and practices during crises or even during the ordinary conflicts of political life.

In light of these circumstances, it is quite possible that the extraordinary pace of democratization begun in the 20th century will not continue long into the 21st century. Nevertheless, the odds are great that in the foreseeable future a very large share of the world's population, in a very large share of the world's countries, will live under democratic forms of government that continue to evolve in order to meet challenges both old and new.

BIBLIOGRAPHY

Democratic institutions. A concise introduction is ALAN F. HATTERSLEY, *A Short History of Democracy* (1930). Historical and theoretical approaches are combined in JOHN DUNN (ed.), *Democracy: The Unfinished Journey, 508 BC to AD 1993* (1992, reprinted with corrections 1993); and SANFORD LAKOFF, *Democracy: History, Theory, and Practice* (1996). General works on ancient Greece include I.E.S. EDWARDS *et al.* (eds.), *The Cambridge Ancient History*, 3rd ed., 14 vol. (1970–2000); and THOMAS R. MARTIN, *Ancient Greece: From Prehistoric Times to Hellenistic Times*, updated ed. (2000). A.H.M. JONES, *Athenian Democracy* (1957, reissued 1986), is indispensable. The most comprehensive study of democracy in Athens is MOGENS HERMAN HANSEN, *The Athenian Democracy in the Age of Demosthenes*, trans. from the Danish by J.A. CROOK (1991, reissued 1999). A brief account of Rome's republican government is F.E. ADCOCK, *Roman Political Ideas and Practice* (1959, reissued 1972). An excellent, though critical, account of the Italian city-state republics is LAURO MARTINES, *Power and Imagination: City-States in Renaissance Italy* (1979, reissued 2002). An essential source on the development of cabinet government in Britain is ARCHIBALD S. FOORD, *His Majesty's Opposition, 1714–1830* (1964, reissued 1979).

The theory of democracy. The theory, foundations, and institutions of democracy are described in ROBERT A. DAHL, *Democracy and Its Critics* (1989, reissued 1991), and *On Democracy* (1998, reissued 2001); and IAN SHAPIRO, *Democracy's Place* (1996).

Problems and challenges. Contemporary problems and challenges are discussed in IAN SHAPIRO and CASIANO HACKER-CORDÓN (eds.), *Democracy's Edges* (1999); KEITH DOWDING, JAMES HUGHES, and HELEN MARGETTS (eds.), *Challenges to Democracy* (2001); and SERGIO FABBRINI (ed.), *Nation, Federalism and Democracy* (2001). Some implications of democratic ideas for nongovernmental organizations are examined in ROBERT A. DAHL, *A Preface to Economic Democracy* (1985); and IAN SHAPIRO, *Democratic Justice* (1999, reissued 2001).

(R.A.Da.)

Denmark

Located strategically at the mouth of the Baltic Sea, the Kingdom of Denmark (Kongeriget Danmark) occupies the peninsula of Jutland (Jylland) and an archipelago of more than 400 islands. Of the country's total land area of 16,639 square miles (43,094 square kilometres), the largest part, 11,497 square miles, is Jutland; the largest of the islands (excluding distant Greenland) are Zealand (Sjælland; 2,876 square miles) and Funen (Fyn; 1,152 square miles). The nation's capital, Copenhagen (København), is located on Zealand; the second largest city, Århus, is the major urban centre of Jutland.

Denmark is attached directly to continental Europe at Jutland's 42-mile (68-kilometre) boundary with Germany. Other than this connection, all the frontiers of Denmark with surrounding nations are maritime, including that with Great Britain to the west across the North Sea. Norway and Sweden lie to the north, separated from Denmark by sea lanes linking the North Sea to the Baltic Sea by way of passages called (from west to east) the Skagerrak, the Kattegat, and The Sound (Øresund). Eastward in the Baltic Sea lies the Danish island of Bornholm.

Though small in territory and population, Denmark has nonetheless played a notable role in European history. In prehistoric times, Danes and other Scandinavians reconfigured European society when the Vikings undertook marauding, trading, and colonizing expeditions. During the Middle Ages, the Danish crown dominated northwestern Europe through the power of the Kalmar Union. In later centuries, shaped by geographic conditions favouring maritime industries, Denmark established trading alliances throughout northern and western Europe and beyond, particularly with Great Britain and the United States. As an important contribution to world culture, Denmark developed humane governmental institutions and cooperative, nonviolent approaches to problem solving.

The Kingdom of Denmark is more than just the land of the Danes. Two remote island worlds in the Atlantic Ocean became integral parts of the Danish state when their colonial status was transformed by full incorporation into the Danish nation. One is the Faroe (Faeroe) Islands, which support a distinctive language and culture. The most remote part of the kingdom is Greenland, an 840,000-square-mile Arctic wilderness, mostly covered by ice, that is the ancestral homeland of scattered coastal communities of Inuit-speaking Greenlanders (also known as Inuit or Eskimos) who formerly lived by hunting and fishing. Many contemporary inhabitants of Greenland are of mixed Danish and aboriginal ancestry. Home rule was granted to the Faroes in 1948 and to Greenland in 1979, though foreign policy and defense remain under Danish control. Each area is distinctive in history, language, and culture.

This article covers the land and people of continental Denmark. For a discussion of its dependent states, see the *Micropædia* articles GREENLAND and FAROE ISLANDS.

The article is divided into the following sections:

Physical and human geography 227
- The land 227
  - Relief
  - Drainage and soils
  - Climate
  - Plant and animal life
  - Settlement patterns
- The people 230
  - Ethnic composition
  - Linguistic composition
  - Religions
  - Demographic trends
- The economy 231
  - Resources
  - Agriculture and fishing
  - Industry
  - Finance and trade
  - Transportation
- Administration and social conditions 232
  - Government
  - Justice
  - Education
  - Health and welfare
- Cultural life 233
  - Daily life
  - The arts and sciences
  - Cultural institutions
  - Recreation
  - Press and broadcasting

History 234
- Earliest settlements 234
- The Viking Age 234
- The 12th, 13th, and 14th centuries 234
  - Kingdom of the Valdemars
  - Dissolution and consolidation
  - Holstein rule and reunion
- The Kalmar Union (1397–1523) 235
  - Rule of Margaret and her heirs
  - The estates
  - The first Oldenburgs
  - The council and the people
  - Civil war and the Lutheran Reformation
- The 17th and 18th centuries 237
  - Introduction of absolutism
  - Foreign policy
  - Trade
- The Napoleonic wars and the 19th century 239
  - The loss of Norway
  - Economic development and the liberal reform movement
  - The National Liberals and the Schleswig-Holstein question
  - The conservative regime
- The 20th century 241
  - Foreign policy and World War I
  - The interwar period
  - Denmark during World War II
  - The postwar period

Bibliography 243

Physical and human geography

THE LAND

The basic contours of the Danish landscape were shaped at the end of the Pleistocene Epoch by the last glaciation of the Ice Age, the so-called Weichsel glaciation. This great glacial mass withdrew temporarily during several warmer interstadial periods, but it repeatedly returned to cover the land until it retreated to the Arctic north for the last time about 10,000 years ago. As a result, the barren layers of chalk and limestone that earlier constituted the land surface acquired a covering of soil that built up as the Weichsel retreated, forming low, hilly moraines that diversify the otherwise flat landscape.

Relief. Denmark proper is a lowland area that lies, on average, not more than 100 feet (30 metres) above sea level. The country's highest point, reaching only 568 feet (173 metres), is Yding Forest Hill in east-central Jutland.

A scenic boundary representing the extreme limit reached by the Scandinavian and Baltic ice sheets runs from Nissum Fjord on the western coast of Jutland eastward toward Viborg, from there swinging sharply south down the spine


Cities over 100,000
Cities 25,000 to 100,000
Cities under 25,000
National capitals
County capitals
RIBE County names
International boundaries
County boundaries
Bridges
National parks
Spot elevations in metres (1 m = 3.28 ft)
Skagerrak
NORTH SEA
Kattegat
BALTIC SEA
Bornholmsgat
SWEDEN
GERMANY
GERMANY
Göteborg
Helsingborg
Copenhagen
Malmö
Aalborg
Århus
Odense
Esbjerg
Frederikshavn
Hirtshals
Skagen
Grenen
Thisted
Hjørring
Brønderslev
Aabybro
Vodskov
Løgstør
Nibe
Klarup
Gistrup
Sæby
Læsø
Ålbæk Bay
Jammer Bay
Tannis Bay
Nissum Fjord
Limfjorden
Nissum Bredning
Løgstør Bredning
Mariager Fjord
Ringkøbing Fjord
Ålborg Bay
Hanstholm
Bal Hill 47
Nykøbing MORS
Sallingsund
Sallingsund Bridge
Struer
Skive
Lemvig
Hobro
Hadsund
Aars
HIMMERLAND PENINSULA
Sjelling Peninsula
Anholt
Viborg
Randers
Grenå
Djursland Peninsula
Holstebro
Bjerringbro
Hadsten
Hinnerup
Hornslet
Løgten
Lystrup
Kjellerup
Hammel
Ringkøbing
Ringkøbing Fjord
Skjern
Virksund
Sunds
Herning
Ikast
Silkeborg
Svejbæk
Gjern
Tilst
Aarhus
Skanderborg
Stilling
Tranbjerg
Hvide Sande
Nymindegab
Tarm
Brande
Farm
Give
Tørring
Horsens
Tvind
Tvind Forest Hill
Tipperne Nature Reserve
Vejle
Jelling
Hedensted
Grindsted
Billund
Varde
Ribe
Esbjerg
Brørup
Vejen
Kolding
Vamdrup
Fredericia
Bogense
Middelfart
Munkebo
Otterup
Kerteminde
Odense
Nyborg
Assens
Faaborg
Bogense
Svendborg
Ærøskøbing
Marstal
Tønning
Rudkøbing
Tårup
Nakskov
Maribo
Sakskøbing
Nykøbing
Gedser
Rødbyhavn
Lolland
Falster
Langeland
Ærø
Fyn
Funen
Sønderjylland
Fanø Island
Rømø Island
Mandø Island
Fehmarn
Fehmarn Belt
Storstrøm
Møn
Møn Cliff
Stevns Cliff
Sjælland Peninsula
Sjælland
Zealand
Vestsjælland
Roskilde
Roskilde Fjord
Isefjord
Holbæk
Kalundborg
Nykøbing
Hundested
Frederikssund
Jægerspris
Frederiksværk
Hillerød
Gilleleje
Hornbæk
Helsingør
Hellebæk
Frederiksborg
Fredensborg
Slangerup
Stenløse
Ballerup
Lyngby
Taarbæk
Hørsholm
Kokkedal
Helsinge
Frederiksberg
Gentofte
Gladsaxe
Herlev
Glostrup
Brøndby
Hvidovre
Rødovre
Tårnby
Dragør
Amager
Køge
Ringsted
Slagelse
Sorø
Korsør
Skælskør
Næstved
Faxe
Præstø
Vordingborg
Stege
Saltholm
Flakfortet
Drogden
Flinterenden
Saltholm Flak
Middelgrunden
Kastrup
Kastrup Bredning
Dragør Sydfælled
Dragør Nordfælled
Kongelige Dyrehave
Bornholm
Rønne
Nexø
Scale 1:2,232,000
1 inch equals approx 35 miles
0 10 20 30 40 mi
0 10 20 30 40 50 60 km
Cone Projection
© Encyclopædia Britannica Inc.

MAP INDEX

Political subdivisions

Århus 56 10 N 10 14 E
Bornholm 55 10 N 15 00 E
Frederiksborg 55 56 N 12 18 E
Fyn 55 20 N 10 25 E
København 55 40 N 12 10 E
Nordjylland 57 00 N 9 50 E
Ribe 55 35 N 8 45 E
Ringkøbing 56 10 N 8 45 E
Roskilde 55 36 N 12 05 E
Sønderjylland 55 15 N 9 15 E
Storstrøm 55 00 N 12 00 E
Vejle 55 45 N 9 20 E
Vestsjælland 55 40 N 11 30 E
Viborg 56 30 N 9 30 E

Cities and towns

Åbenrå (Aabenraa) 55 02 N 9 26 E
Åbybro 57 09 N 9 45 E
Ålborg (Aalborg) 57 03 N 9 56 E
Århus (Aarhus) 56 09 N 10 13 E
Års 56 48 N 9 32 E
Assens 55 16 N 9 55 E
Augustenborg 54 57 N 9 53 E
Beder 56 04 N 10 13 E
Bellinge 55 20 N 10 20 E
Billund 55 44 N 9 07 E
Bjerringbro 56 23 N 9 40 E
Bogense 55 34 N 10 06 E
Brædstrup 55 58 N 9 37 E
Bramdrupdam 55 31 N 9 28 E
Bramminge (Bramming) 55 28 N 8 42 E
Brande 55 57 N 9 07 E
Brønderslev 57 16 N 9 58 E
Brørup 55 29 N 9 01 E
Copenhagen (København) 55 40 N 12 35 E
Dianalund 55 32 N 11 30 E
Dragør 55 36 N 12 41 E
Erritsø 55 33 N 9 42 E
Esbjerg 55 28 N 8 27 E
Fåborg 55 06 N 10 15 E
Fakse 55 15 N 12 08 E
Fensmark 55 17 N 11 49 E
Fredericia 55 35 N 9 46 E
Frederiksberg 55 41 N 12 32 E
Frederikshavn 57 26 N 10 32 E
Frederikssund 55 50 N 12 04 E
Frederiksværk 55 58 N 12 02 E
Galten 56 21 N 10 03 E
Gilleleje 56 07 N 12 19 E
Gistrup 57 00 N 10 00 E
Give 55 51 N 9 15 E
Gråsten 54 55 N 9 36 E
Grenå 56 25 N 10 53 E
Grindsted 55 45 N 8 56 E
Haderslev 55 15 N 9 30 E
Hadsten 56 20 N 10 03 E
Hadsund 56 43 N 10 07 E
Hammel 56 15 N 9 52 E
Haslev 55 20 N 11 58 E
Hedensted 55 46 N 9 42 E
Hellebæk 56 04 N 12 34 E
Helsinge 56 01 N 12 12 E
Helsingør 56 02 N 12 37 E
Herning 56 08 N 8 59 E
Hillerød 55 56 N 12 19 E
Hinnerup 56 16 N 10 04 E
Hirtshals 57 35 N 9 58 E
Hjørring 57 28 N 9 59 E
Hobro 56 38 N 9 48 E
Højby 55 20 N 10 27 E
Holbæk 55 43 N 11 43 E
Holstebro 56 21 N 8 38 E
Høng 55 31 N 11 18 E
Hornbæk 56 05 N 12 28 E
Hørning 56 05 N 10 03 E
Hornslet 56 19 N 10 20 E
Horsens 55 52 N 9 52 E
Hundested 55 58 N 11 52 E
Hvalsø 55 36 N 11 52 E
Hvide Sande 55 59 N 8 08 E
Ikast 56 08 N 9 10 E
Ishøj 55 37 N 12 19 E
Jægerspris 55 51 N 11 59 E
Jyderup 55 40 N 11 26 E
Jyllinge 55 45 N 12 07 E
Kalundborg 55 41 N 11 06 E
Kerteminde 55 27 N 10 40 E
Kjellerup 56 17 N 9 26 E
Klarup 57 01 N 10 03 E
København, see Copenhagen
Køge 55 27 N 12 11 E
Kolding 55 29 N 9 29 E
Kolt 56 06 N 10 04 E
Korsør 55 20 N 11 09 E
Langeskov 55 22 N 10 36 E
Lemvig 56 32 N 8 18 E
Lind 56 06 N 8 59 E
Løgstør 56 58 N 9 15 E
Løgten 56 17 N 10 19 E
Løgumkloster 55 03 N 8 57 E
Løsning 55 48 N 9 42 E
Lystrup 56 14 N 10 15 E
Maribo 54 46 N 11 31 E
Middelfart 55 30 N 9 45 E
Munkebo 55 27 N 10 34 E
Næstved 55 14 N 11 46 E
Nakskov 54 50 N 11 09 E
Neder Holluf 55 21 N 10 27 E
Neksø 55 04 N 15 09 E
Nibe 56 59 N 9 38 E
Nordborg 55 03 N 9 45 E
Nyborg 55 19 N 10 48 E
Nykøbing 54 46 N 11 53 E
Nykøbing 55 55 N 11 41 E
Nykøbing 56 48 N 8 52 E
Odder 55 58 N 10 10 E
Odense 55 24 N 10 23 E
Ølgod 55 49 N 8 37 E
Ølstykke 55 47 N 12 11 E
Otterup 55 31 N 10 24 E
Padborg 54 49 N 9 22 E
Præstø 55 07 N 12 03 E
Randers 56 28 N 10 03 E
Ribe 55 21 N 8 46 E
Ringe 55 14 N 10 29 E
Ringkøbing 56 05 N 8 15 E
Ringsted 55 27 N 11 49 E
Rødekro 55 04 N 9 21 E
Rønne 55 06 N 14 42 E
Roskilde 55 39 N 12 05 E
Rudkøbing 54 56 N 10 43 E
Ry 56 05 N 9 46 E
Sæby 57 20 N 10 32 E
Sakskøbing 54 48 N 11 39 E
Silkeborg 56 10 N 9 34 E
Skælskør 55 15 N 11 19 E
Skagen 57 44 N 10 36 E
Skanderborg 56 02 N 9 56 E
Skive 56 34 N 9 02 E
Skjern 55 57 N 8 30 E
Slagelse 55 24 N 11 22 E
Slangerup 55 51 N 12 11 E
Snejbjerg 56 08 N 8 55 E
Sønderborg 54 55 N 9 47 E
Sorø 55 26 N 11 34 E
Stege 54 59 N 12 18 E
Stenløse 55 46 N 12 12 E
Stilling 56 04 N 10 00 E
Støvring 56 53 N 9 51 E
Strib 55 32 N 9 47 E
Struer 56 29 N 8 37 E
Sunds 56 12 N 9 01 E
Svejbæk 56 08 N 9 38 E
Svendborg 55 03 N 10 37 E
Svogerslev 55 38 N 12 01 E
Tarm 55 55 N 8 32 E
Thisted 56 57 N 8 42 E
Thurø 55 03 N 10 40 E
Tilst 56 12 N 10 07 E
Toftlund 55 11 N 9 04 E
Tønder 54 56 N 8 54 E
Tørring 55 51 N 9 29 E
Tranbjerg 56 06 N 10 09 E
Vamdrup 55 25 N 9 17 E
Varde 55 38 N 8 29 E
Vejen 55 29 N 9 09 E
Vejle 55 42 N 9 32 E
Viborg 56 26 N 9 24 E
Videbæk 56 05 N 8 38 E
Vodskov 57 06 N 10 02 E
Vojens 55 15 N 9 19 E
Vordingborg 55 01 N 11 55 E

Physical features and points of interest

Ær Island 54 53 N 10 20 E
Ålbæk Bay 57 35 N 10 30 E
Ålborg Bay 56 45 N 10 30 E
Als, *island* 54 59 N 9 55 E
Amager, *island* 55 37 N 12 37 E
Anholt, *island* 56 42 N 11 34 E
Århus Bay 56 09 N 10 18 E
Arnå, *river* 54 57 N 8 53 E
Arre, Lake 55 58 N 12 08 E
Baltic Sea 55 30 N 15 00 E
Bornholmsgat, *marine channel* 55 20 N 14 30 E
Bornholm, *island* 55 10 N 15 00 E
Bul Hill 57 09 N 9 02 E
Djursland Peninsula 56 20 N 10 45 E
Endelave, *island* 55 46 N 10 17 E
Esrum, Lake 56 00 N 12 24 E
Fakse Bay 55 10 N 12 15 E
Falster, *island* 54 48 N 11 58 E
Fan Island 55 25 N 8 25 E
Fej Island 54 57 N 11 26 E
Fem Island 54 58 N 11 33 E
Frø Bavne Hill 55 20 N 10 07 E
Funen (Fyn), *island* 55 20 N 10 30 E
Fur, *island* 56 50 N 9 02 E
Great Belt, see Store Strait
Grenen, *spit* 57 44 N 10 40 E
Gudenå, *river* 56 29 N 10 13 E
Gyldenlove Hill 55 33 N 11 52 E
Hessel Island 56 12 N 11 43 E
Himmerland Peninsula 56 50 N 9 45 E
Jammer Bay 57 20 N 9 30 E
Jutland (Jylland), *region* 56 00 N 9 15 E
Kattegat, *strait* 57 00 N 11 00 E
Knøsen Hill 57 12 N 10 18 E
Køge Bay 55 30 N 12 20 E
Læs Island 57 16 N 11 01 E
Langeland, *island* 55 00 N 10 50 E
Langeland Strait 54 50 N 10 55 E
Lille Strait 55 20 N 9 45 E
Lim Fjord, *channel* 56 55 N 9 10 E
Lolland, *island* 54 46 N 11 30 E
Man Island 55 16 N 8 34 E
Møn, *island* 55 00 N 12 20 E
Møn Cliff 54 58 N 12 33 E
Mors, *island* 56 50 N 8 45 E
Nissum Bay 56 38 N 8 22 E
Nissum Fjord, *lagoon* 56 21 N 8 14 E
North Sea 56 45 N 7 50 E
Odense, *river* 55 26 N 10 26 E
Omme, *river* 55 55 N 8 25 E
Øresund, see Sound, the
Randers Fjord, *bay* 56 36 N 10 20 E
Ringkøbing Fjord, *lagoon* 56 00 N 8 15 E
Røm Island 55 08 N 8 31 E
Rye, *river* 57 06 N 9 47 E
Salling Peninsula 56 40 N 9 00 E
Sams Island 55 52 N 10 37 E
Sejer Island 55 53 N 11 09 E
Silkeborg Lakes 56 05 N 9 34 E
Sjælland, see Zealand
Sjælland Peninsula 55 58 N 11 22 E
Skagerrak, *strait* 57 45 N 9 00 E
Skjern, *river* 55 55 N 8 24 E
Småland Sound 55 05 N 11 20 E
Sound (Øresund), the 55 50 N 12 40 E
Stevn Cliff 55 18 N 12 27 E
Storå, *river* 56 19 N 8 19 E
Store Strait (Great Belt) 55 30 N 11 00 E
Tannis Bay 57 40 N 10 15 E
Tipperne Nature Reserve 55 54 N 8 13 E
Tis Lake 55 35 N 11 18 E
Varde, *river* 55 35 N 8 20 E
Yding Forest Hill 56 00 N 9 48 E
Zealand (Sjælland), *island* 55 30 N 11 45 E

The ice front

of the peninsula toward Åbenrå and the German city of Flensburg, just beyond the Danish frontier. The ice front is clearly marked in the contrast between the flat western Jutland region, composed of sands and gravels strewn by meltwaters that poured west from the shrinking ice sheet, and the fertile loam plains and hills of eastern and northern Denmark, which become markedly sandier toward the prehistoric ice front.

In northern Jutland, where the long Lim Fjord separates the northern tip from the rest of the peninsula, there are numerous flat areas of sand and gravel, some of which became stagnant bogs. Burials and ritual deposits interred in these bogs in antiquity—especially during the Bronze and Iron ages—have been recovered by archeologists. In more recent centuries, these bogs were a valued source of peat for fuel. In the 20th century, they have been drained to serve as grazing areas for livestock.

In places along the northern and southwestern coasts of Jutland, salt marshes were formed by evaporation of an inland sea that existed during the Late Permian epoch (approximately 258 to 245 million years ago). Senonian chalk, deposited about 100 million years ago, is exposed in southeastern Zealand, at the base of Stevn Cliff and Møn Cliff, and at Bulbjerg, in northwestern Jutland. Younger Danian limestone (about 65 million years old) is extensively quarried in southeastern Zealand.

On Bornholm, outcroppings reveal close affinities with geologic formations in southern Sweden. Precambrian granites more than 570 million years old—among the oldest on the Earth's surface—are exposed across extensive areas on the northern half of the island. On the southern half, Cambrian sandstone and shales overlie the older granites.

Drainage and soils. The longest river in Denmark is the Gudenå. It flows a distance of 98 miles from its source just northwest of Tørring, in east-central Jutland, through the Silkeborg Lakes and then northeast to empty in the Randers Fjord on the east coast. There are many small lakes; the largest is Arre on Zealand, with a surface area of 15.7 square miles. Large lagoons have formed behind

the coastal dunes in the west, such as at the Ringkøbing and Nissum fjords.

In most of Denmark, the soil rests on glacially deposited gravel, sand, and clay, under which lie ancient chalk and limestone. The subterranean limestone resulted in permeation of the soil with calcium that diminished its value for agriculture when it was first brought under cultivation in the Neolithic era. Through millennia of cultivation, however, the soil improved greatly, so that 60 percent of the land surface is excellent for farming.

Climate. Denmark experiences changeable weather because it is located in the temperate zone at the meeting point of diverse air masses from the Atlantic, the Arctic, and eastern Europe. The west coast faces the inhospitable North Sea, but the terminal section of the warm Gulf Stream (the North Atlantic Drift) moderates the climate. The cold, rainy winters produce frozen lakes and snow-covered fields, yet they are moderated by the influence of the Gulf Stream. The mean temperature in February, the coldest month, is 32° F (0° C), which is 12° F (7° C) higher than the worldwide average for that latitude. The number of days when freezing weather occurs ranges annually from 70 on the west coast to 120 in the interior. Summers are mild, featuring episodes of cloudy weather interrupted by sunny days. The mean temperature in July, which is the warmest month, is 61° F (16° C).

Influence of the Gulf Stream

Rain falls throughout the year but is relatively light in winter and spring and greatest from late summer to early winter. The annual precipitation of 25 inches (635 millimetres) ranges from about 32 inches in southwestern Jutland to about 16 inches in parts of the archipelago.

Plant and animal life. In prehistoric times, before fields were cleared for cultivation, much of the land was covered with a deciduous forest of oak, elm, lime (linden), and beech trees. The original forest did not survive, but highly valued areas were reforested later to break up the expanses of agricultural fields that dominate the landscape. Denmark borders the coniferous belt and has therefore been receptive to the establishment of plantations of spruce and fir, particularly in parts of Jutland where extensive wastelands of dune vegetation and heather were reclaimed for forestry. In all, about 10 percent of the land is forested.

Abundant postglacial herds of large mammals, including elk, aurochs, brown bear, and wild boar, died out under the pressures of human expansion and an intensive agricultural system. Roe deer, however, still occupy the countryside, and the large-antlered red deer can be found in the forests of Jutland. Smaller animals such as hares and hedgehogs have also survived. Birds are abundant, numbering more than 300 species, of which about half breed in the country. Storks—common summer residents in the early 20th century—migrate each year from their winter home in Africa, but they are now rare. Fish are abundant in Danish waters, particularly cod, herring, and plaice, which form the basis for a large fishing industry.

Settlement patterns. Agriculturalists established a village settlement pattern early in the prehistory of Denmark. From at least the Middle Ages until the 18th century, these settlements were organized under the rules of an open-field system, the dominant feature of which was communalism. Most individual landholders were tenant farmers (*fæstebønder*) whose farm buildings and land belonged to the local manor house (*herregård*). The scattered plots of tenanted land were located in each of two or three large fields, which were farmed collectively by the tenants; therefore, it was essential that villagers agree on the nature and timing of plowing, harrowing, planting, and harvesting. Meeting at a central place in the village, family heads discussed common problems of field management and agreed on mutual responsibilities and cooperation. Each family received harvests from its own plots but worked with the others to manage the fields. They shared resources in order to assemble large wheeled plows, each drawn by six or eight horses. Livestock were grazed as a single village herd on the stubble of harvested fields. Shared decisions were also made on the use of communal facilities such as the meadow, commons, village square, pond, and church. Danish peasants cooperated in much of what they did, and a communal spirit was the product.

Open-field system



Farms surrounding the town of Nørreby, Fem Island, Denmark.
Erik Beiting/Pressehuset

The open-field system was replaced by the consolidation of fields (*udskiftningen*) and the purchase of farms (*frikøbet*) as a result of the great land reforms (*De store landboreformer*) put into place by reform-minded estate owners. By the beginning of the 19th century, the wheeled plow had been replaced by a lightweight plow that could be pulled by a single horse, which most farmers could afford. The bulk of the economy shifted from subsistence to commercial farming. The result was dismantlement of the old open-field system and an end to village communalism. As small scattered plots were consolidated into larger individual holdings, some landowners moved their farmsteads away from the village to be closer to their fields, obscuring the pattern of village settlements. Parts of western Jutland were late to be brought under cultivation. Poor soil resulting from postglacial geologic conditions discouraged farmers from settling in these areas until population pressure made them more attractive.

The nation is no longer overwhelmingly rural. An economic shift to light industry and trade was associated with a growth in the size of towns and cities. In the late 20th century the movement to cities continued but was no longer primarily to major centres. Migration to smaller urban centres grew disproportionately, and migrants to urban areas settled more commonly in suburban residential communities than in the cities as such. With modern roads and railway transportation, in a land in which distances are not great, the isolation of farms and villages has ended.

THE PEOPLE

Ethnic composition. Denmark is almost entirely inhabited by ethnic Danes. Very few Faeroese or Greenlanders have settled in continental Denmark, despite their status as Danish citizens. Small German, Jewish, and Polish minorities, on the other hand, have been long established and are substantially assimilated. In the 1960s an economic expansion required more labour than the nation could supply, and "guest workers" (*gæstearbejdere*) made their way into Denmark. In the late 1980s the most numerous ethnic minorities in Denmark were Turks, Yugoslavs, Iranians, and Pakistanis.

"Guest workers"

Linguistic composition. Danish is the official language. It is closely related to Norwegian, with which it is mutually intelligible, especially in the written form. Although the other Scandinavian languages (excepting Finnish) are close relatives, they are sufficiently different to be understood easily only by those schooled or experienced in the effort. Many educated or urban Danes have learned to speak a second language. English has replaced German as the most popular foreign language.

Religions. Religious freedom is an unchallenged value

in Denmark. Roman Catholic churches and Jewish synagogues have long existed in the larger cities. About 90 percent of all Danes, however, are at least nominally members of the state Evangelical Lutheran church (*folkekirken*), which replaced Roman Catholicism as the official religion after the Lutheran Reformation in 1536.

"Grundtvigianism" designates a revitalization movement that inspired a new sense of Christian awareness in the 19th century at a time when Danish Protestantism had become very formal and ritualistic. N.F.S. Grundtvig provided a philosophical, religious, and organizational basis for "educating and awakening" the impoverished peasantry. This was achieved by establishing folk high schools in which Christian belief and peasant culture were taught as a basis for creating pride in the Danish heritage.

The Home Mission

A revival movement within the framework of the Danish church was also organized in the 19th century. Known as the Home Mission (Indremissionen), it was founded by a clergyman, Vilhelm Beck. The Home Mission survives as a contemporary evangelical expression of Pietism, which had won converts in the 18th century. Members of the Home Mission constitute a minority within the church; they place emphasis on the importance of individual Bible study, personal faith, and a sin-free style of living.

Demographic trends. For many years the net population total has just barely maintained itself. The total fertility rate (average number of births for each childbearing woman) is only 1.4. A nearly universal acceptance of the concept of population control, in combination with emigration to the United States and elsewhere, has effectively restrained growth. Because losses in reproductive replacement and emigration have been offset by slight increases from immigration, the population remained nearly stable at 5.1 million during the late 20th century. The age distribution has shifted as a consequence of a lower level of fertility. There are now relatively fewer persons in the under-20 age group and more in the over-80 age group.

THE ECONOMY

Denmark supports a high standard of living with well-developed social services. It boasts a per capita gross national product that is one of the highest in the world. The economy is based primarily on service industries, trade, and manufacturing; only 6 percent of the population is engaged in agriculture, fishing, and forestry. Small enterprises are dominant.

The only Nordic country to do so, Denmark joined the European Economic Community in 1973, at the same time as the United Kingdom, then its most important trading partner. At the same time, economic collaboration among the Nordic countries continues. No passports are required for travel by Scandinavians within the region, and communication among the various agencies of government is direct and need not be channeled through their respective embassies. Scandinavians enjoy a common labour market that includes reciprocal social welfare benefits and the right to vote in local elections in the neighbouring country of residence. There is capital mobility, supported by the Nordic Investment Bank. Uniform legislation, particularly with regard to commercial law, dates to the 19th century.

In the Danish mixed welfare-state economy, private sector expenditures account for more than half of the net national income. Public expenditure is directed to education, national defense, social services, and agricultural subsidies. The government neither owns capital nor has significant commercial or industrial income. Public income is primarily derived from taxes on real estate, personal income, and capital and through customs and excise duties. The heaviest indirect tax, which goes to the national government, is the value-added tax (VAT).

Both employers and employees are well organized. Membership in unions is normally based upon the particular skills of the workers. The association of employees is called the National Confederation of Trade Unions (Landsorganisationen); the principal association of employers is the Danish Confederation of Employers (Dansk Arbejdsgiverforeningen).

Resources. Danish natural resources are quite limited. During the early 1970s the economy suffered from dependence on imported petroleum for more than 90 percent of its energy needs. Finds of oil and natural gas fields in the Danish sector of the North Sea permitted a partial self-sufficiency in this regard. Coal-fired power plants produce 90 percent of the nation's electricity, up from 10 percent in 1970. The switch from petroleum was accompanied by economies of production in which otherwise wasted heat from the production of electricity is used to heat water that is piped to homes and factories. By this means, the energy output of power plants has been doubled.

Agriculture and fishing. A new spirit of communalism among farmers emerged with the massive social changes at the end of the 19th century, a time of poverty and economic depression. When cereal prices fell, Danish farmers fought for survival by using their crops as fodder to produce butter, eggs, and bacon. They succeeded by establishing folk high schools as well as agricultural and dairy cooperatives. The result was a peasantry that was literate, well motivated, and competitive in the marketplace. The producer cooperatives disbanded after 1950, however, and folk high schools have lost their central importance.

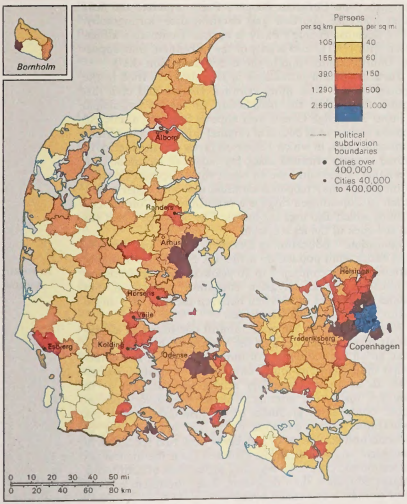
Major crops

Next to its well-educated labour force, the soil is still Denmark's most important raw material. About 60 percent of the land is intensively exploited and extensively fertilized. More than half of the cultivated land is devoted to cereals, with barley and wheat accounting for a large percentage of the total grain harvest. Sugar beets are another leading crop. Oats, rye, turnips, and potatoes are grown in western Jutland, where the soil is less fertile.

Domesticated animals are an important feature of life in Denmark. Dairy cattle, pigs, and poultry are raised in great numbers to supply both the domestic and the foreign markets.

Farms are generally small or medium-sized family-owned enterprises. The extensive fertilization and application of scientific animal husbandry helps to maintain the viability of small farm operations. Milk and dairy products, pork, and eggs account for a major percentage of the total value of production. Fur farming, especially of minks and foxes, is economically important.

The fishing industry is still economically important. Her-



Population density of Denmark.

ring, cod, and plaice (or flatfish) account for more than 90 percent of the total catch; other important species include salmon, eel, and deepwater shrimp. Danish commercial fishing also extends into foreign waters in search of Atlantic cod, Norwegian pout, and North Sea sprat (brisling).

Industry. Denmark has a small extractive industry that relies on granite (for roads and housing) and kaolin (for ceramics and paper manufacture) found on Bornholm. Local boulder clays are molded and baked to make bricks and tiles. Moler (marine diatomaceous earth) is mined for use in insulating materials for the building industry, and white chalk is essential for the manufacture of cement.

The largest employers are the manufacturers of metal products, machinery, and equipment; the food processing industry; the paper and graphic industries; and manufacturers of transport equipment. The production of footwear, clothing, wood and wood products, furniture, and electronic equipment also provide substantial employment.

Finance and trade. In 1846, the first commercial bank was established in Denmark. In 1975, commercial and savings banks became equal in status, and foreign banks, which theretofore had maintained representative offices in Copenhagen, were permitted to establish full branches. All banks are under government supervision, and public representation is required on their supervisory boards.

The banking system

The National Bank of Denmark (Danmarks Nationalbanken) is the only bank of issue and enjoys a special status as a self-governing institution under government supervision. Profits revert to the state treasury. The national stock exchange, established in 1861, is located in Copenhagen.

Imports of raw materials and fuel formerly were balanced largely by exports of agricultural products, supplemented by income from shipping and tourism. In the late 20th century the overseas trade pattern shifted to a major reliance on the export of industrial products, including industrial machinery, electronic equipment, and chemical products. Agricultural products such as fish, poultry, dairy products, meat, beer, and furs are still important exports, however. Denmark has created an export market for household furniture, toys, silverware, ceramics, plastics, textiles, clothing, and other goods notable for their creative modern design.

As a member of the EEC (since 1993, the European Community [EC]), Denmark relies heavily upon foreign trade within Europe. Germany, the United Kingdom, and Sweden provide the largest markets for both agricultural products and manufactured goods.

Transportation. An extensive road and highway system serves the nation. The number of private automobiles in use has risen rapidly since World War II. Bicycles, once a common mode of transport, are still popular. Cities and towns maintain bicycle lanes located parallel to motor roads and sidewalks.

Bus and coach routes extend throughout the country, including some operated by the Danish State Railways and others organized regionally by local government authorities. A comparatively large railroad network was established during the last half of the 19th century.

Characteristic features of the Danish transportation system are the ferries and many bridges. Of particular importance are the bridge and tunnel systems that connect Zealand with Funen (via the small island of Sprogø) and Copenhagen with Malmö, Swed. (opened 1997–98 and 2000, respectively). Several bridges also connect Funen and Jutland. Many good harbours provide favourable conditions for both domestic and international shipping.

Kastrup, near Copenhagen, is one of the busiest airports in Europe. In addition to internal flights, the airport at Kastrup is a centre for international air traffic. The Scandinavian Airlines System (SAS), a joint Danish-Norwegian-Swedish enterprise, flies European and intercontinental routes. Danair, owned by SAS, operates regular services between Copenhagen and other cities on Zealand, Jutland, Funen, Bornholm, and the Faroe Islands.

ADMINISTRATION AND SOCIAL CONDITIONS

Government. The constitution of June 5, 1953, provides for a unicameral legislature, the Folketing, with 179 members (including two from the Faroe Islands and two from Greenland). The prime minister heads the government, which is composed additionally of cabinet ministers who run the various departments, such as justice, finance, and agriculture. The monarch signs acts passed by the Folketing upon the recommendation of the cabinet sitting as a Council of State. To become law, the acts must also be countersigned by at least one cabinet member. Faced with a vote of no confidence, the cabinet must resign.

The constitution of 1953

In addition to establishing unicameralism, the 1953 constitution mandates popular referenda (used, for example, to secure public approval for Danish entry into the EEC) and postulates the creation of an ombudsman office—the first outside Sweden, its country of origin. The Succession to the Throne Act, which accompanied the 1953 constitution, provides for female succession. This allowed the accession of Queen Margrethe II in 1972.

Denmark is divided into 13 counties and 271 municipalities. It has universal adult suffrage by voluntary and secret ballot, with a voting age of 18 for both national and local elections. All voters are eligible to run for office. The voter turnout in national elections approaches 90 percent. Elections are held on the basis of proportional representation, in which each political party gains seats in the Folketing, or in the city council, in proportion to its strength among the voters. As a result the national government is usually composed of a coalition of parties that does not enjoy a majority, and the government must piece together a majority for each item of legislation. Members of the Folketing are elected to a four-year term, but the prime minister may dissolve the legislature and call for new elections at any time. Despite the splintering of parties, Denmark has enjoyed stable government, with new elections on an average of once every three years.

Political parties

The largest Danish political party, the Social Democratic Party, led most Danish governments from the 1930s to the early 1980s; since then, coalition governments have predominated, including those of nonsocialist parties headed by the Conservative People's Party and the Liberal Party in 1981–93. All of Denmark's political parties support the continuation of the welfare state, except for the tax-protest Progress Party, which also expresses anti-immigration sentiments. The Christian People's Party criticizes the laws that liberalize abortion and decriminalize pornography. The Socialist People's Party is against Denmark's affiliation with NATO and is part of the movement that opposes Danish membership in the European Union (EU).

Justice. Minor infractions in Denmark are tried in the police courts. Most other criminal charges and civil disputes fall within the jurisdiction of the 84 municipal courts. Two High Courts hear appeals from the municipal courts and serve as courts of original jurisdiction in serious criminal cases, in which 12-person juries are impaneled. In some nonjury criminal cases, lay judges sit alongside professional judges and have an equal vote. A special Court of Complaints may reopen a criminal case and order a new trial. In Copenhagen there is a Maritime and Commercial Court, which also uses lay judges. The Supreme Court sits at the apex of the legal system.

Education. Education in Denmark is free, and virtually the entire adult population is literate. Nine years of school attendance for children 7 to 16 years of age is compulsory. Preschool and kindergarten education is optional, but about 60 percent of children between the ages of 5 and 6 years attend kindergarten classes.

After reaching the 9th grade, almost one-fourth of all students leave school to enter the workforce, approximately half go on for additional vocational education, and somewhat more than one-fourth continue on to matriculate either in an upper secondary school (*gymnasium*) or in another institution offering a higher preparatory education. While many graduates of the gymnasium enter the workforce, others continue on to the university or to schools and academies of university rank that specialize in technical and artistic fields. Children from the highest socioeconomic class have a greater probability of continuing higher education than do others. The enrollment of men and women is essentially equal.

At the pinnacle of higher education are the University of

Higher education

Copenhagen (founded in 1479), the University of Aarhus (1928), and the University of Odense (1964), all state-supported. In addition, university centres were established at Roskilde in 1970 and Ålborg in 1974.

Health and welfare. Danes on the whole enjoy excellent health. Aggressive public health programs are directed against the threats of infectious diseases. Public health nurses provide free advice and assistance to mothers, which, with good nutrition and housing, has contributed to a low infant mortality rate. More than 80 percent of the cost of the health care system is paid for by national and local authorities and employers. Life expectancy at birth is about 72 years for males and about 78 years for females.

Danish citizens may choose between two primary health care options. Most Danes opt for completely free care that is provided by a general practitioner; some, however, prefer to pay about one-third of their medical bills out of pocket for the privilege of choosing any family physician or specialist they wish.

A national old-age pension scheme is available for all persons 67 years of age and older, with basic amounts being paid irrespective of the financial position of the beneficiary. A disability pension and a pension for widows 55 years of age or older are also in effect.

According to the Danish constitution, "Any person unable to support himself or his dependants shall, where no other person is responsible for his or their maintenance, be entitled to receive public assistance." The state welfare programs of Denmark should not be thought of as institutionalized charity, however. Health, education, unemployment, disability, and old-age benefits are available at virtually no charge to all Danes. They are recognized both legally and in public opinion as morally just social rights that have been paid for by taxes and assessments.

CULTURAL LIFE

Daily life. Danes traditionally faced life from the security of the nuclear family, as has been true throughout Europe. During the late 20th century, substantial changes have taken place. For example, marriage is no longer entered into by young adults as an almost inevitable social institution. Historically, the Danes easily tolerated sexual relations between individuals who were engaged to be married. In earlier centuries it was not uncommon for marriage to take place after a baby was born, although it was considered immoral and unacceptable not to marry eventually. Now, the inevitability of marriage has fallen away. Cohabitation without the formalities of engagement and wedding is common. Nearly one-fifth of all unions in Denmark are by cohabitation rather than formal marriage. Consistent with the decline of contracted marriages, the incidence of divorce has risen. One marriage in four may be expected to end in dissolution.

Forty percent of live births now take place out of wedlock, as compared with only 10 percent a generation ago. These children are not necessarily raised by single parents, however. Children are born to approximately 40 percent of consensual unions, and two children or more are found in 15 percent of such relationships. The changes in marriage and divorce statistics and the growing incidence of consensual unions are primarily due to the changed role of women in society. Women have experienced greater independence as well as increased responsibility for economic survival and child care. They are educated on a more equal basis with men, and they participate more equally in the job market, although not yet with equal pay. The availability of contraceptive methods and free abortions has also increased women's options. In the mid-1960s slightly fewer than 50 percent of married women between the ages of 20 and 50 engaged in paid employment. Twenty years later more than 80 percent of married women were working. The ability to earn their own incomes has made marriage less necessary for women to provide security for themselves and their children. It has also made divorce less punitive in socioeconomic terms.

Status of women

The arts and sciences. Although the Danes are few in population, they have been numerous in contributing to the growth of world civilization. Tycho Brahe (1546–1601) was a major figure in the early telescopic exploration of the universe; Thomas Bartholin (1616–80) was the first anatomist to describe the human lymphatic system; Nicolaus Steno (1638–86) established geology as a science; Ole Rømer (1644–1710) measured the speed of light for the first time; Caspar Thomeson Bartholin, Jr. (1655–1738) discovered the ductus sublingualis major and the glandula vestibularis major, both of which bear his name as Bartholin's duct and gland; Hans Christian Ørsted (1777–1851) discovered electromagnetism; Niels Finsen (1860–1904) won the Nobel Prize for Physiology or Medicine for his work on the medical uses of ultraviolet rays, and Johanes Fibiger (1867–1928) won the same award for his research on cancer; Valdemar Poulsen (1869–1942) developed a device for generating radio waves; Niels Bohr (1885–1962) won the Nobel Prize for Physics for his achievements in quantum physics, a prize which was later won by his son, Aage Bohr; and Carl Peter Henrik Dam (1895–1976) won the Nobel Prize for Physiology or Medicine for the discovery of vitamin K.

Saxo Grammaticus (d. 1204) contributed a book of history, *Gesta Danorum*, to world literature; Rasmus Rask (1787–1832) founded comparative philology; N.F.S. Grundtvig (1783–1872) founded a theological movement and pioneered in education relating to human rights; Søren Kierkegaard (1813–55) helped to shape existentialist philosophy; Bertel Thorvaldsen (1770?–1844) achieved renown as a sculptor in a neoclassic style; Hans Christian Andersen (1805–75) authored fairy tales that are read throughout the world; Carl Nielsen (1865–1931) composed classical music of international fame; Carl T. Dreyer (1889–1968), a film director, is respected internationally; Jørn Utzon won world recognition as the architect of the Sydney Opera House; and Karen Blixen (1885–1962) achieved world acclaim writing under the name of Isak Dinesen. The Nobel Prize for Literature was awarded to the novelist Henrik Pontoppidan (1857–1943) in 1917 and to Johannes V. Jensen (1873–1950), whose works included the novel *The Long Journey* (*Den lange rejse*), in 1944.

Cultural institutions. The first Danish-speaking theatre was opened in Copenhagen in 1722, followed in 1748 by The Royal Theatre (*Det Kongelige Teater*), which remained under court patronage for a century. In 1848 it was taken over by the state and is now administered by the Ministry for Cultural Affairs. Besides a relatively large number of classical and modern Danish plays, the repertoire includes much that is current in Britain, the United States, Germany, and France.

Ballet and theatre

A resident ballet company, which also performs in The Royal Theatre, was founded more than 200 years ago, but only through its youngest generation of dancers in the style of choreographer August Bournonville has it become internationally acclaimed as the Royal Danish Ballet.

Denmark supports 10 symphony orchestras; two of the more important are the Danish Radio Symphony Orchestra and the Royal Orchestra. Musicians and singers are trained at the Royal Danish Conservatory in Copenhagen and other conservatories and at the Opera Academy.

The Royal Danish Academy of Fine Arts was established in 1754. It produced Bertel Thorvaldsen and, in the 20th century, the sculptor Robert Jacobsen and the architects Arne Jacobsen and Kaj Gottlob. Famous craft concerns include the firm of Georg Jensen, silversmith, the Royal Copenhagen and Bing and Grøndahl porcelain factories, the glassworks of Holmegård and Kastrup, and the furniture manufacturer Fritz Hansens Eftf.

Recreation. The pursuit of sport became popular after defeat in the Danish-German war of 1863–64 as Danes turned to an interest in small arms and physical training. Soon every part of Denmark had established shooting, gymnastics, and athletic clubs. Soccer was introduced to Denmark by British engineers who came to design the railroad system in the 1870s. Rowing was organized at a national level as early as 1886, making the Copenhagen Rowing Club (established in 1865) the oldest sports club in the world. Soccer, or association football, became an organized sport when the Copenhagen Ball Club was established in 1876, and soccer remains a national sport. At various times Danish athletes have won Olympic gold medals in canoeing, shooting, swimming, rowing, and

cycling events. These and many other sports appeal to sports-minded Danes, particularly in the summer months, and no Dane is immune to the attractions of a visit to one of the many well-tended parks, forests, or beaches that honeycomb the nation.

Press and broadcasting. Radio Denmark offers Danish programming, but in most parts of Denmark it is also possible to receive strong radio signals from neighbouring nations, particularly Sweden in the north and Germany in the south. State-managed television broadcasting is mainly devoted to current affairs, cultural events, and programs of interest to children and young people. The owners of radio and television sets pay an annual license fee, which finances broadcasting operations and frees the main radio and television stations from the interruptions of commercial advertising. A second television channel and some radio transmission is funded partly by commercials.

Complete freedom of the press is guaranteed under the constitution. About 50 newspapers under private ownership are published throughout the nation. Most are associated with a political party. The largest dailies are *Ekstrabladet, BT, Berlingske Tidende,* and *Politiken.*

(R.T.A./S.V.A.)

For statistical data on the land and people of Denmark, see the *Britannica World Data* section in the BRITANNICA BOOK OF THE YEAR.

History

The first written evidence of a Danish kingdom dates from the early Viking Age. Roman knowledge of this remote country was fragmentary and unreliable, and the traditional accounts in *Widsith* and *Beowulf* and by later Scandinavian writers, notably Saxo Grammaticus (*c.* 1200), are too mythical and legendary to serve as history. Since World War II, however, archaeological research and the study of place-names have provided considerable information about the earliest settlements.

EARLIEST SETTLEMENTS

The first nomadic hunters, after 12,000 BC, developed a Stone Age culture. About 4000 BC one of the greatest changes in Danish history occurred: the inhabitants adopted the practice of agriculture and stock keeping, and the first farmers began to reclaim land from the forests. From about 3500 BC permanent houses for the dead in large megalithic graves were built, but about 2800 BC a single-grave culture emerged. The change was caused by local factors, including new tools, weapons, and religious rites. In the last phase of the Stone Age, the Dagger period (2400–1800 BC), flint working reached its apogee with the production of technical masterpieces, including daggers and spearheads that were imitations of imported metal weapons.

Bronze and Iron ages

The refined culture of the ruling class in the Bronze Age (1800–500 BC) is indicated by the spiral decorations on the bronzes of the period, in particular the famous Late Bronze Age *lurs* (long, curved metal horns, often found in pairs) from about 1000–800 BC. At about the same time, the wooden plow enabled better exploitation of the cultivated areas.

After 500 BC, bronze was gradually replaced by iron, and a village society developed in a landscape of bogs, meadows, and woods with large clearings. The villages appear to have been moved and the fields abandoned with each new generation. Chiefs and rich farmers lived in houses between 40 and 100 feet in length, the climate now being colder and wetter; as in the Bronze Age, objects of great value were laid as offerings in the bogs. The period up to AD 400 was marked by the large number of villages, and splendidly equipped graves suggest that political power was gathered in fewer hands. More or less fixed trading connections were established with the Romans; about AD 200 the first runic inscription appeared, possibly developed under the influence of the Latin alphabet. The period from 400 to 800 is known as the Germanic Iron Age, but the finds have been few, indicating a time of decline, unrest, and bubonic plague in the 6th century. The first trading markets appeared at Hedeby (near what is now Schleswig, Ger.) and Ribe in the 8th century, and written sources mention the existence of slaves. (M.I.A.L.)

THE VIKING AGE

The northward expansion of the Franks brought Denmark into close contact with European powers. To protect the country from military aggression from the south, a great rampart, the Dannevirke (Danewirk), was built along the border from the Baltic to the North Sea, near the modern town of Schleswig. Dendrochronological dating has shown that the wall was erected shortly after 737, which seems to indicate the formation of a state at that early period. Later, in 808, the Frankish annals describe the building of a wall by King Godfrey (d. 810) and campaigns against other Danish kings. In about 960 the Dannevirke was connected with the wall around Hedeby, the largest city; other centres were Roskilde (on Zealand) and Jelling, probably the seats of the first kings.

Louis I the Pious, the son of Charlemagne, tried to Christianize the Danes. Louis sent a monk, Ansgar, to Hedeby in 826, but his message was resisted; later, however, he was given permission to erect churches in Hedeby and Ribe. The following year he was installed as bishop of Hamburg with the whole of Scandinavia as his see. After Ansgar's death, in 865, his successor, Rimbert, wrote a hagiographic account of his life, *Vita Ansgarii,* which is an important source for 9th-century Scandinavian history.

During the 10th century, after internal struggles between rival kings, the centre of power moved to Jelling, where Gorm became king of Jutland (*c.* 940). On a huge runestone at Jelling, his son, Harald I Bluetooth (Blåtand), attributed to himself the unification of all Denmark, the conquest of Norway, and the Christianization of the Danes. It is possible that he agreed to become a Christian (*c.* 960) in order to avoid German meddling in Denmark, although he was later forced to protect the southern border from a German attack. Probably Harald started the building of the great circular fortresses of Trelleborg (Zealand), Nonnebakken (Funen), and Fyrkat and Aggersborg (Jutland), dated about 980. Under the king's protection the new bishops of Jutland Christianized the kingdom, and the gravesite of Harald's pagan parents at Jelling was made into a Christian shrine. Harald's conquest of Norway was short-lived, and his son Sweyn I Forkbeard and grandson Canute I the Great were each forced to rewin the country. Sweyn exhausted England in annual raids and formed an Anglo-Danish kingdom, a policy that his son and grandson continued until the latter's death in 1042. English missionaries were sent into Denmark to counteract the power of the Hamburg archbishops, but Denmark remained within the orbit of the German prelates. Norway elected a native king in 1035, who also ruled Denmark from 1042 to 1047, when Canute's nephew Sweyn II Estridson was chosen as king. During his reign, in the 1070s, Adam of Bremen composed his *Gesta Hammaburgensis ecclesiae pontificum* (*History of the Archbishops of Hamburg-Bremen*), the first important contemporary source for Danish history.

Unification under Harald Bluetooth

THE 12TH, 13TH, AND 14TH CENTURIES

Working together with the church, Sweyn Estridson strengthened royal power. The country was divided into eight bishoprics: Slesvig, Ribe, Århus, Viborg, Vendsyssel, Odense, Roskilde, and Lund. The royal succession remained in the hands of the local *things.* Five of Sweyn's sons succeeded each other on the throne—Harald Hén (ruled 1074–80), Canute II the Holy (Knud; 1080–86), Oluf Hunger (1086–95), Erik Ejegod (1095–1103), and Niels (1104–34). Their reigns were marked by popular and ecclesiastical opposition to the extent of royal power, as, for instance, during the reign of Canute, whose bailiffs were ruthless in their treatment of the peasants. A rebellion in Vendsyssel forced the king to flee to Odense, where he was killed in St. Alban's Church. Under Erik Ejegod, Scandinavia was recognized in 1103 as an archbishopric with a see at Lund (Skåne), where a great Romanesque cathedral was erected.

In order to defend the southern border, Niels made Erik Ejegod's son Knud Lavard duke of South Jutland. Knud's success against the Wends on the south coast of the Baltic

won him great popularity but also the ill will of the king and his supporters, in particular Niels's son Magnus the Strong, who killed Knud in 1131, causing a civil war. Knud's brother Erik Emune took up the fight against Magnus. In 1134 Erik's army defeated that of Magnus, who was killed along with 5 bishops and 60 priests. Shortly after the battle, Niels visited Slesvig (Schleswig), a centre of Knud Lavard's support, and was killed by the townspeople; his successor was Erik Emune (1134–37). The civil war continued, and by 1146 the kingdom was divided between the sons of Erik Emune, Magnus the Strong, and Knud Lavard. After continued struggles, Knud's son Valdemar was acknowledged as the sole king in 1157.

Kingdom of the Valdemars. The Wends continued their attacks on the Baltic trade and the Danish coast; when the Germans increased their expansion eastward along the Baltic coastline, Valdemar I the Great (1157–82) allied himself with the Saxon prince Henry III the Lion against the Wends and acknowledged the Hohenstaufen emperor Frederick I Barbarossa as his overlord. With the blessing of the church, represented by Absalon, the bishop of Roskilde (later archbishop of Lund), Valdemar undertook repeated crusades against the Wends; in 1169 he captured Rügen, placed the island under his rule, and began to establish Danish hegemony over the Baltic, as described by Saxo Grammaticus. The cooperation of king and church resulted in the crowning of Valdemar's son Canute IV (also called Canute VI) as king in 1170 by the archbishop Eskil; a vigorous state was established and German claims of overlordship were rejected.

The reign of Valdemar

In the early 1180s north Germany was split among petty counts, and Absalon, who ruled during Canute's minority, attacked Pomerania and annexed it along with part of Mecklenburg to the Danish realm (1184). Canute's brother Valdemar, count of South Jutland, defeated the count of Holstein, adding the county to his own territory. When Valdemar II, called Sejr (Victor), became king (1202), the land between the Elbe and the Eider, including Lübeck, was brought under the Danish crown. Valdemar conquered Estonia in 1219, and, according to legend, Dannebrogen, the national flag, came down from heaven during the siege of Revel (Tallinn), which became a strong fortress, marking the culmination of Danish rule over the Baltic. In 1223 Valdemar was taken prisoner by one of his north German vassals but bought his freedom in 1225, promising to give up all the conquered areas except Estonia and Rügen. A final attempt to win back the lost areas led to his decisive defeat in 1227.

Valdemar's son Erik was crowned (1232) during his father's lifetime, and his other sons, Abel and Christopher, were proclaimed dukes; Abel was given South Jutland and Christopher received the islands of Lolland and Falster. As a check against royal misuse of power, a parliament, the *hof*, was established by the high prelates and aristocrats; it met at short intervals and also functioned as the highest court. During Valdemar's reign two essential works appeared: a code of law and King Valdemar's *Jordebog* ("Land Book"; a cadastre, or land register).

Dissolution and consolidation. Soon after Valdemar's death in 1241, a struggle broke out between his sons; Erik was killed in 1250 by the forces of his brother Abel, who succeeded him but soon lost his life during a war on the Frisians in 1252. Christopher was then crowned king, and Abel's eldest son, also called Erik, became duke of South Jutland, which was soon after declared a hereditary duchy. Under Christopher I, the cooperation between church and crown ended. The archbishop, Jakob Erlandsen, demanded the full extension of canon law, but was opposed by both the king and the peasants. Erlandsen was taken prisoner by the king, and Denmark was placed under an interdict. The archbishop was supported by Erik, duke of South Jutland, the count of Holstein, and the prince of Rügen, who attacked Denmark. During the ensuing war Christopher died (1259).

The regents for Christopher's young son, Erik V Glipping, released the archbishop, who left the country. When Erik became king, he was forced by the *hof* to sign a coronation charter (1282), in which he agreed to assemble the *hof* each year, to have no one imprisoned purely on suspicion, and to respect "King Valdemar's Law"; this was the first written constitution. In 1286 Erik was murdered; the election of his son Erik Menved as king (1286–1319) without a charter was a sign that the importance of the *hof* was declining. Erik Menved tried to take advantage of the weakness of the north German states, but he was unable to maintain his early gains because of the country's financial inability to support a mercenary army: at his death in 1319 the state finances were chaotic.

Development of the *hof*

The childless Erik Menved was succeeded by his brother, Christopher II, who was forced to sign a strict charter and was the first king to accept the *hof* as a permanent institution, independent of his personal supporters. He did not abide by the charter, however, and he was driven into exile after a battle with the magnates and the count of Holstein. For a time (1326–30), the young duke of South Jutland, Valdemar, ruled under the regency of the count of Holstein. After Christopher's return, the kingdom was split by a peasant uprising, church discord, and the struggle with the Holsteiners, who received almost all of the country in pawn; Skåne rebelled against its Holstein count and came under Swedish rule.

Holstein rule and reunion. The counts of Holstein ruled Denmark from 1332 to 1340, when one of them was murdered during a visit to Jutland, and Christopher's son, Valdemar IV Atterdag, was chosen king. He married the sister of the duke of South Jutland, who gave the northern quarter of North Jutland as her dowry; he began his reign with the reunion of Denmark as his first priority. By selling Estonia (1346) and collecting extra taxes, he reclaimed some of the pawned areas and brought others back through negotiations or force of arms. In 1360 he conquered Skåne and, a year later, Gotland, and Denmark was reunited. During his reign royal power was strengthened. At a *hof* in 1360, a "great national peace" was agreed between the king and the people. The *hof* was replaced by the Rigsråd—a national council of the archbishop, the bishops, and *lensmænd* (vassals) from the main castles—and the king's Retterting (Court of Law) became the supreme court. Valdemar also attacked major economic problems: after the Black Death (1349–50), he confiscated ownerless estates and regained royal estates that had been lost during the interregnum; the army was reorganized.

Valdemar's war on Gotland and the fall of the wealthy town of Visby brought him into conflict with Sweden and the Hanseatic League, which declared war on Denmark. Sweden's king, Magnus II Eriksson, agreed to the marriage (1363) of his son, crown prince Haakon (Haakon VI of Norway), to Valdemar's daughter Margaret; however, Magnus was soon overthrown by Swedish magnates and replaced by his nephew Albert of Mecklenburg (1364). In 1367 the Hanseatic League, the princes of Mecklenburg and Holstein, and some of the Jutland magnates attacked Valdemar at sea and on land. The king went to Germany to find allies in the rear of his powerful German enemies and succeeded in obtaining a rather favourable peace treaty at Stralsund in 1370, which gave the Hanse trading rights in Denmark and pawned parts of Skåne to the league for 15 years. Valdemar returned home and continued his work of stabilizing the crown's hold on the country. After his death in 1375 the magnates elected Olaf, the five-year-old son of Margaret and Haakon, on condition that he signed a charter. His father died in 1380, and Olaf, under Margaret's regency, also became king of Norway and called himself heir to the Swedish throne.

THE KALMAR UNION (1397–1523)

Rule of Margaret and her heirs. Denmark, Norway, and Sweden were united during the 14th century by dynastic ties. Margaret, who served as regent of both Denmark and Norway during Olaf's minority, worked to win the crown of Sweden for him, but he died in 1387. She was acknowledged as regent in the two countries, and in 1388 rebellious Swedish nobles, dissatisfied with Albert's rule, hailed her as regent in Sweden. The following year her troops defeated Albert's knights and captured the king. The war continued until 1398, when Stockholm was finally turned over to Margaret, who wanted to secure the royal succession; in June 1397 her sister's grandson, Erik

of Pomerania, was crowned king of Denmark, Norway, and Sweden at Kalmar. There negotiations with the Scandinavian magnates were initiated to regulate the mutual relations of the three countries and the power of the new king. There is no record of the discussions, only the coronation charter on parchment and the union document on paper. The latter concerns the conditions for the union, including the election of kings and the internal matters of each country (that is, each would be governed by its own laws). The union document is among the most controversial sources for Scandinavian history: present opinion considers it a draft, signed by only 10 of the 17 magnates mentioned and never accepted by the queen or her successors.

The union document

The primary objective of Queen Margaret, who ruled the union until her death in 1412, was to strengthen royal power. In Denmark she avoided calling the national council, left high posts unoccupied, and reduced the privileges of the nobles. Her economic policy was the most successful aspect of her reign: with the help of extra taxes, she managed to pay back loans, reclaim pawned areas, buy Gotland from the powerful Teutonic Knights, who had occupied the island in 1398, and donate land and gifts to the church. Margaret's foreign policy was based on a desire to keep peace, and she arranged for the marriage of her successor, Erik, to Princess Philippa, daughter of Henry IV of England.

Erik attacked the Hansa in 1410, and after his enthronement in 1412 he promoted national trade in Denmark and privileges for English merchants; he also gave Danish towns a monopoly on commerce and crafts (1422) and began to collect custom duties in the Sound in 1429 to replace the lost revenues from the Skåne market. His foreign policy was not successful: the conflict with the Hansa was intensified, and he also made enemies of the Teutonic Knights. The king increased the number of Danes appointed to offices in Sweden and Norway, arousing the anger of the native aristocracies, and his efforts to control church appointments irritated the clergy. Constant warfare made heavy demands on the exchequer and forced him to increase taxation and debase the coinage, which worsened the economic situation. In 1413 he had the Danish *hof* recall the fiefs of the Holsteiners, which started a long war, and in 1426 Lübeck and its allies opened hostilities over trade privileges and blockaded the Scandinavian countries, which especially affected the mining districts of Sweden. In 1434 a rebellion broke out in Bergslagen (central Sweden) under Engelbrekt Engelbrektsson, who was elected guardian of the realm. He and the Swedish council renounced their allegiance to Erik. The spirit of revolt spread to Erik's enemies in Denmark and Norway, and in 1438 he went into exile in Gotland. He was deposed in all three kingdoms during the years 1438–42.

Erik's nephew, Christopher of Bavaria, was called by the Danish council to become king in 1440 and was later accepted as king in both Sweden and Norway. He promised to administer the three countries separately and to use only native *lensmænd*. When Christopher died without heirs in 1448, the union was temporarily dissolved and the Danish council elected Count Christian of Oldenburg king as Christian I; in the following year he was also proclaimed king of Norway. (H.En./M.I.A.L.)

Until 1520 the union was marked by wars between Denmark and Sweden, interrupted by periods of peace. The Danish kings and most of their nobles sought to follow a policy of supremacy, which was opposed by the Swedish kings and guardians of the realm. Throughout the period the struggle for power between the king and the magnates continued, although some Swedish nobles supported the kings of Denmark.

The estates. During this period the people became more sharply divided into estates. Agriculture was then, as at all times, the principal industry; the cultivated land, apart from about 1,000 manors, consisted of about 80,000 farms, clustered together in groups of 5 to 20 as villages. These were managed by peasant farmers in common, whether they owned their farms themselves or were tenants paying a yearly rent (*landgilde*). In 1500 about 12,000 peasants owned farms, about 18,000 were leasehold tenants of crown lands, and about 30,000 were leasehold tenants of lands belonging to the church or the nobles.

Peasants, church, and nobility

With its seven bishoprics and more than 70 monasteries, the church was immensely rich. It derived a huge income from its lands and farms and drew still greater revenues from the tithes on the entire grain production of the country, one-third going to the bishops, one-third to the parish churches, and one-third to the parish priests. Since the Council of Basel, the pope had assumed the right to make all ecclesiastical appointments, although he allowed certain nominations by the king. The nobles, however, tried to reserve some of these for their younger sons, who were too poor to buy manors.

The 15th century marks a turning point in the history of the Danish nobility. Until then any Dane could become a noble by presenting himself well-equipped for military service at his own expense. In return he was exempted from all taxes; but from the 15th century he had to show that his forefathers had enjoyed tax exemptions for at least three generations. The king sought to assume the right to issue titles of nobility, but despite this the nobility in this period developed the characteristics of a caste. During the 15th century the nobility comprised 264 families, but this number fell to 230 in 1500 and to 140 (including at most 3,000 persons) in 1650; the Gyldenstjerne and the Rosenkrantz (whose names are commemorated in Shakespeare's *Hamlet*) were among the most important.

The nobility acquired lands in great numbers and were capable agriculturists, responsible for increased exports of farm produce. The country had a long-standing market for its horses; now stall-fed bullocks were added. Landowners, lay and clerical, also became merchants, many of them having their own ships. (A.E.Cn./M.I.A.L.)

The first Oldenburgs. The first three kings of this still (through the collateral branch of Glücksborg) reigning house—Christian I (1448–81), John (Hans; 1481–1513), and Christian II (1513–23)—tried to foster the economies of the towns while curbing the direct trade of German Hanseatic merchants with the peasants.

Christian I and the coronation charter

With Christian I began the revival of the coronation charter, which now included a guarantee for the national council's participation in foreign policy, legislation, taxation, and justice. If the king failed to respect the guarantees, his subjects had the right to renounce him, but this did not prevent the early Oldenburgs from ignoring the charters. Christian attempted to circumvent the council by calling a meeting of the estates (1468), a practice followed by his successors. After the death of the last male heir to the Holstein counts, Christian reached an agreement with the family whereby he became count of Slesvig (South Jutland) and duke of Holstein (1460), with the two areas to be "eternally undivided" and ruled by a royal heir chosen by the local nobility. In paying off a number of princes who had claims on the territories, Christian ran heavily into debt, which, together with the costs of the war with Sweden, forced him to pawn the Norwegian Shetland and Orkney islands to the Scottish king in order to provide a dowry for his daughter.

One of the major concerns of the first Oldenburgs was to reestablish the Kalmar Union. Danish opinion held that the country's power depended on the union, and many Swedish nobles desired a union so long as their influence on home affairs could be maintained. In 1464, however, a rebellion broke out that Christian's troops failed to suppress, and his attempt in 1471 to force the Swedes back into the union was unsuccessful.

Christian died in 1481 and was succeeded by his son John (also king of Norway from 1483), who wanted to reduce the power of the nobility and the Hansa and to create a strong Nordic monarchy with support from the peasants and burghers. Many administrative posts were given to nonnobles, and the king signed trade agreements with the Dutch and English. He was also acknowledged as Swedish king in 1483, but he was not crowned until 1497, after a war between the two countries. He ruled in Sweden until 1501, when rebellious Swedish nobles recalled Sten Sture the Elder, who had served as guardian of the realm from 1470 to 1497. Another war between Denmark and Sweden lasted from 1506 to 1513, and from 1510 to 1512

John was also involved in a successful war with Lübeck.

Christian II succeeded his father in 1513. A struggle broke out in Sweden between the Sture party (led by Sten Sture the Younger) and the union party (led by Archbishop Gustav Trolle); in 1520, after Trolle had been captured by the Stures, a Danish army attacked Sweden and defeated the Sture army. In November 1520 Christian II was crowned hereditary king of Sweden, and 82 members of Sture's party were executed in the "Stockholm Bloodbath." Christian returned to Denmark, leaving the Swedish government in the hands of the archbishop and his allies, who soon faced a rebellion led by Gustav Vasa. After a period of warfare Vasa was elected regent of Sweden (1521). In Denmark Christian attempted to increase his power by ignoring the nobles and replacing them with men from the burgher class; he also interfered with the affairs of the church. The opposition to the king grew, and in 1523 the members of the council from Jutland renounced allegiance to him and joined his enemies. Lübeck and Christian's uncle, Frederick of Holstein-Gottorp, joined the rebellion, and he was crowned king of Denmark and Norway as Frederick I in 1523, when Christian II fled from the country. Sweden elected Gustav I Vasa, and the Kalmar Union was permanently dissolved.

The "Stockholm Bloodbath"

The council and the people. The peasantry suffered a decline under the Kalmar Union. The towns enticed young people from the farms, which, together with a reduction of the labour force due to the Black Death, caused an increase in abandoned farms. This led to semi-serfdom, *vornedskab,* practiced especially in Zealand. By the 16th century, those tenured peasants who lived near the manor worked off a portion of their taxes by service at the manor.

Under the Kalmar Union, Danish towns prospered and the influence of the burghers grew. By 1500 there were approximately 80 towns, most of them fortified but all small; Copenhagen had at most 10,000 inhabitants. The monopoly on internal trade granted by Erik of Pomerania improved the economic position of the burghers, and many German merchants took out citizenship in the towns in order to compete.

Disappearance of the *hof*

During the Kalmar Union, the old *hof* disappeared and the participation of the provincial *things* in legislation and royal elections ceased, while the people were represented by the estates, which remained unimportant throughout the period. The national council, composed of the bishops and nobles chosen by the king, including the highest civil servants (the *hofmester,* or master of the court, in charge of finances; the *drost,* the chief political and judicial officer; the *kansler,* or chancellor; the *marsk,* or marshall, in charge of military forces; and the *rigsadmiral,* or admiral), held the power of legislation and taxation together with the king. The council's consent was necessary for declarations of war, and, together with the king, it served as the highest court, but the entire council was seldom called, and it had little influence on daily administration. The fiefs were controlled by the king's representatives, who collected taxes and upheld the law; these positions were never made hereditary.

The major political conflict during the Kalmar Union was between the monarchs and the nobility. The high nobility never accepted Christian II's strong monarchy, and its members were his most bitter enemies. Frederick I (1523–33) adopted a cautious policy toward both the nobility and the peasants, and he tried to reconcile the Danish, Dutch, and Hanseatic merchants; thus, when Christian II tried to regain his power by invading Norway in 1531–32, there was no national rising, and Christian was taken prisoner. On accession to the throne, Frederick had promised the bishops to fight heresy. Actually, he invited Lutheran preachers to the country, most probably to expand the royal power at the expense of the church.

Civil war and the Lutheran Reformation. After the death of Frederick I in 1533 the Catholic and conservative majority of the Rigsråd once more triumphed. They postponed the election of a new king, fearing that the obvious candidate, Prince Christian (later King Christian III), if elected, would immediately introduce Lutheranism. They tried unsuccessfully to sponsor his younger brother Hans. Civil war, however, broke out in 1534, when the burgomasters of Malmö and Copenhagen accepted help from the Lübeckers, who, under the pretext of restoring Christian II, hoped to regain their mercantile supremacy and control of the Sound. The landing of Lübeck troops in Zealand in the early summer of 1534 roused the Jutland nobility. Now even the Roman Catholic bishops supported Prince Christian. Count Christopher of Oldenburg was leader of the forces of Lübeck, while Christian's general was the Holstein noble Johan Rantzau, a Lutheran. Rantzau subdued a revolt of the Jutland peasants and the civil war ended in the summer of 1536. The Catholic bishops were taken into custody and their property confiscated; the monasteries were dissolved and vast estates came to the crown.

In October 1536 the estates sanctioned a Danish Lutheran Church and in 1537 appointed new bishops, all of burgher descent. They had, however, no political influence, as bishops no longer sat in the Rigsråd. The church organization was finally established in 1539. The purged national council that emerged after the Reformation was soon able to assert itself. The charter issued by Christian III differed only slightly from earlier ones with regard to the privileges of the nobility and the constitutional power of the Rigsråd. The king's attempt to make the throne hereditary did not quite succeed. The Rigsråd named Prince Frederick as the successor of his father and the king's charter provided that a Danish prince should always be elected, but this was omitted in Frederick II's charter in 1559. The national council thus suffered no permanent loss of elective power.

(As.F./M.I.A.L.)

The strengthened monarchy

The monarchy was decisively strengthened by the civil war, primarily by the confiscation of church property. The nobility no longer had much to gain as an independent political body, and they chose to take part in the politics of the strengthened monarchy. A noble could be a member of the Rigsråd, govern a county on behalf of the king and the council, or simply cultivate his domain to profit from the rising prices on grain and cattle. The merchants of Copenhagen and Malmö had fought Christian III, but they favoured a strong monarchy that would protect their interests in the Baltic trade. The monarchy built a strong public administration in Copenhagen (Chancery, Rent Chamber) and even in far-off Norway. The foundations of absolutism were laid during Christian's peaceful reign.

The strain on the public finances during the reign (1559–88) of Frederick II, resulting from the war with Sweden, was relieved through heavier taxation on the farmers. But the main income came from a duty in the Sound on the constantly increasing trade in the Baltic. Originally a fixed duty per ship, it was changed into a fee on tonnage; it was at the king's own disposal, out of reach of the council. The Sound was considered Danish national waters; this fiction and the Sound duty remained until 1857.

THE 17TH AND 18TH CENTURIES

Christian IV (1588–1648), who succeeded his father at the age of 10, thus had favourable political and economic conditions for his ambitious policies. For seven years, an aristocratic regency headed by the aging chancellor, Niels Kaas, was able to influence the future ruler. The first half of Christian's personal reign was in every respect a success, marked by the dynamic king's many initiatives: establishing trading companies, acquiring overseas possessions, investing in a colony in India at Tranquebar, founding new towns, and erecting monumental buildings in the capital and elsewhere. The strongest incentive in his foreign policy was to secure Danish control of the Baltic, into which Sweden was expanding. Christian reacted by intervening in the Thirty Years' War to strengthen the position of Protestantism and to secure a broad sphere of interest in Germany as a counterweight to Swedish expansion, but he was defeated in 1626. From this reversal dates the gradual decline of Denmark-Norway's role in European politics. Nevertheless, the king's national government, public administration, jurisdiction, and promotion of business and new industries had great importance for the future.

Christian IV is regarded as one of the greatest Danish rulers, a central figure in later drama, poetry, and art. But in reality the military catastrophes weakened the position

of the monarchy, so that the high nobility decided to curtail the power of his successor, Frederick III (1648–70).

Introduction of absolutism. Absolutism was nevertheless introduced during Frederick's reign, when the magnates proved unable to handle a central government. After the military debacles in 1658–60 (when Sweden's Charles X Gustav attacked Jutland from the south and marched to Zealand over the frozen sounds of Funen—the Belts—in the winter of 1658; Denmark lost Skåne, Halland, and Blekinge, and Norway lost Bohuslän), the nobles even refused to pay any taxes. The situation in 1660 was exploited by the king's councillors, who drafted a new constitution that eliminated special political privileges of the nobility and proclaimed the king absolute sovereign. This constitution (and a secret "King's Law" of 1665, which is said to be the most absolutist of all European theories of absolutism) lasted until 1848 with only minor modifications.

Denmark's governing bureaucracy

After 1660 Denmark was governed by an efficient bureaucracy, but the political leaders came from the class of great privileged landowners; wealth, not noble birth, gave access to this class. The government in Copenhagen consisted of "colleges." There were five as a rule—namely, the old Chancery, the Rent Chamber, and new colleges for commerce, war, and the navy. Top decisions were made by a secret council, in which leaders of the colleges could easily influence the king. Local administration remained largely unchanged after 1660, but the government took pains to curtail the military power of the new county governors (*amtmænd*). The absolutist kings, very unlike their Swedish colleagues, were rather anonymous, in part because of their feeble mental powers.

After 1660 the crown reduced its properties, which had been greatly increased by the Reformation, through sales to its bourgeois creditors (who thus moved into the class of great landowners). The state compensated for loss of income by increasing taxes on the land according to the value of the holding of each peasant. The new assessments made in the period from 1660 to the 1680s served as the bases of taxation in both Denmark and Norway until the 19th century. Until 1660 the king and the council had acted as supreme court; in 1661 the Danish supreme court was created, and appeals could be made to it from the whole kingdom. Law was codified in Denmark in 1683.

Foreign policy. Denmark's participation (1709–20) in the Great Northern War demonstrated that even with alliances it had no hope for recapturing the territories it had lost to Sweden during the preceding century. On the other hand, Sweden no longer had the strength to invade Denmark from the south in alliance with the house of Gottorp (Slesvig). The king decided on a careful foreign policy to keep a balance in the north and to safeguard communications between Denmark and Norway. This necessitated alliances with Russia and the Netherlands and, from time to time, France. This policy succeeded for the rest of the 18th century, probably because of the common European need for free access to the Baltic. In the 1770s the Gottorp lands in Schleswig and Holstein were brought under Danish rule.

The importance of freedom of the seas

During the 18th century Denmark-Norway acquired an important merchant marine and a navy. Freedom of the seas had become a vital issue and a difficult problem, complicated especially by the export of Norwegian timber to England. During wars in the middle of the century (1740–63) Denmark-Norway had to bow to the British claim of ruling the waves. But during the U.S. War of Independence (1775–83) the Danish foreign minister Andreas Peter Bernstorff in 1780 organized an armed neutrality treaty with Holland and Sweden, whose King Gustav III had married a Danish princess. Because of the war between England and France, Denmark and Sweden prolonged the treaty in 1794, which Russia and Prussia renewed in 1800. Norwegian export interests would have been threatened, however, if England considered these treaties hostile acts, so in 1780 Bernstorff also concluded a special treaty with England, much to the annoyance of Russia. Such a policy of balance proved to be impossible after 1800. (G.Sa./M.I.A.L.)

Trade. Denmark, poor in natural resources except for its soil, made no important economic gains in the 18th century. No important industry developed. Following mercantilist theory the government supported trade, to the benefit of Copenhagen merchants. But Denmark lacked the political strength to exploit the strategic position of Copenhagen; imports dominated its trade. Except for oxen and meat, Denmark had very little to export. Eastern Norway was made an outlet for Danish grain in the 1730s, but the grain was inferior and normally could not compete with Baltic grain on the western European markets.

The principal reason for Denmark's stagnant economy was the backward state of Danish agriculture in the 18th century. A body of some 300 Danish landlords owned about 90 percent of the Danish soil, grouped in 800–900 estates. The landlords were the real rulers of the country, because their social position gave them privileged positions in filling the leading posts in the administration, the chanceries, and the Rent Chamber.

A price depression beginning in the 1720s enabled the landlords to use their position to impose very strict laws and regulations on the Danish peasants, who lived in villages, renting their farms from landlords whose demesne lands alone covered about 10 percent of the land. To get cheap labour, a compulsory provision that one live in the village of his birth was required for all people between 4 and 40 years of age. As the system was coupled with military conscription, the landlord could threaten a young peasant with at least six years of army service if he did not rent a farm. Every tenant had to perform labour on the landlord's domain for an average of three days a week. This work was considered to be the rent of the peasant's holding. A tenant had no right to demand a contract when he took over a holding, nor could he demand payment for improvements he might make on the holding when the lease expired or was lifted by the landlord. Each landlord also had the right of petty jurisdiction on his estate. Even if the landlords got cheap labour and the army received sufficient manpower by this system, Danish agriculture suffered from incredibly low productivity. The farmers performed poorly on the domains of the landlords; they had too little time to cultivate their own holdings; and they had no reason whatsoever to improve them.

Compulsory service

Until the 1780s, Danish society seemed stagnant. A financial crisis in the 1760s, after Russian threats during the Seven Years' War, was solved by a lasting poll tax. In 1770 a German doctor, Johann Friedrich Struensee, depending on palace intrigues and the queen's bed, gained control of the government through the half-mad king Christian VII (1766–1808). For a year and a half, freedom of the press and intense reform activity reigned, but mostly on paper. Struensee had no popular support, and he naturally provoked resistance and fury among the landlord class. He was arrested for crime against the majesty and, in April 1772, was sentenced to death. His short reign has been classified as another failure of the enlightened despotism of the 18th century. The old men who came back to power led an even more reactionary policy than before, one lasting until a bloodless coup d'état in 1784.

The years between 1784 and 1797 have been called the happiest period in all Danish history. Danish politics of those years were led by Bernstorff, Ditlev Reventlow, and Ernst Schimmelmann, all from the landlord class, by the benevolent crown prince Frederick (Frederick VI, 1808–39), and by the Norwegian jurist Christian Colbjørnsen. Notable reforms included liberal custom tariffs (1797) and the abolition of the Danish grain monopoly in Norway. But above all, the time was an age of land reforms, beginning in 1786 and lasting until the state suffered financial bankruptcy in 1813. Sixty percent of the Danish peasants became landowners. Compulsory residence, compulsory labour on the domains of landlords, and private jurisdiction were abolished, and the land was redistributed and made into independent farms. The army had to get soldiers by ordinary conscription. Landlords were compensated for the rights they lost, and together with the new landowning farmers they were assured stable labour by strict legislation on the landless crofters. The land reforms were possible because of a continuous rise in grain prices between 1750 and 1815 and because the new men of 1784 had carried out successful reforms on their own estates. As responsible

Agricultural reforms

politicians they had an insight into the benefits of a mild inflation and a liberal allocation of state credit, with which they guided the transition to peasant landownership. No doubt the revolution in France, beginning in 1789, influenced this evolution. The example of the independent French farmers after 1789 hindered an evolution like that in England, with great domains and a numerous class of rural workers. From the land reforms, and from the school act of 1814, which introduced compulsory schooling for all children between ages 7 and 14, there stemmed a high standard of Danish agriculture. The Danish land reforms are remarkable also as the only successful feat of European enlightened despotism.

THE NAPOLEONIC WARS AND THE 19TH CENTURY

The Napoleonic Wars of the early 19th century tore Denmark-Norway out of a peaceful period that had lasted since 1720. The armed neutrality treaty of 1794 between Denmark and Sweden, to which Russia and Prussia adhered in 1800, was considered a hostile act by England. In 1801 a detachment of the English navy entered The Sound and destroyed much of the Danish fleet in a battle in the harbour of Copenhagen. When the English fleet next proceeded to threaten the Swedish naval port of Karlskrona, Russia started negotiations with England. The result was a compromise, which Sweden was forced to adopt in 1802. The neutrality treaty had fallen in ruins. Denmark-Norway, nevertheless, managed to keep out of the wars and to profit from them until 1807.

The Treaty of Tilsit (1807) between France and Russia worsened the situation. In 1805 France had lost its fleet to the English at the Battle of Trafalgar. England feared that the continental powers might force Denmark to join them so that the Danish navy could be used to invade England. To eliminate this threat, England resorted not to diplomacy but to force. In September English troops occupied Zealand and an English fleet bombarded Copenhagen. Denmark had no choice but to capitulate to the English demands. On October 20 the English commander sailed away with the whole Danish fleet. The "fleet robbery" was severely criticized, even in the British House of Commons. Because of fear of a French or Russian occupation, Denmark chose what seemed to be the lesser evil and joined the continental alliance against England on October 31. This step also meant war with Sweden. Denmark might have reacted differently if England had used diplomacy, but the events of September had been too much of an affront to the Danish government and especially to the crown prince Frederick. An alliance with England was no longer possible.

The loss of Norway. The continental blockade of England, which was against Danish interests, was a catastrophe to Norway. Fish and timber exports were stopped, as well as grain imports from Denmark. The consequences were isolation, economic crisis, and hunger. In 1810–13 England consented to some relaxation of its counterblockade against Norway. As a whole, however, the years 1807–14 convinced leading groups in Norway that they needed a political representation of their own.

Denmark-Norway remained an ally of Napoleon until 1814. After Napoleon's defeat at the Battle of Leipzig (1813), Sweden repeated its 17th-century strategy by attacking Denmark from the south. By the Treaty of Kiel (Jan. 14, 1814), Denmark gave up all its rights to Norway (but not to the old Norwegian dependencies of Iceland, the Faroes, and Greenland) to Sweden.

Sweden's attack from the south

The Danes did not intend this agreement to end their union with Norway. While remaining outwardly loyal to the Treaty of Kiel, the Danish government worked for the eventual return of Norway. This probably is why the crown prince Christian Frederick, governor of Norway, in collusion with the Danish king, organized an uprising against the Treaty of Kiel. A constituent assembly was called by Christian Frederick to meet at Eidsvoll, 40 miles north of Christiania (modern Oslo, Nor.). It drew up a constitution (which still exists) on May 17, 1814, and elected Christian Frederick to the throne of Norway.

Norwegian independence got no support from the great powers, and Sweden attacked Norway in July 1814. After a fake war of 14 days, Christian Frederick resigned, and Danish hopes of reunion were lost. (G.Sa.)

Economic development and the liberal reform movement. *Economic consequences of the war.* The Napoleonic Wars had proved a national catastrophe for Denmark, both economically and politically. The policy of armed neutrality had failed, and that part of the fleet not destroyed had been surrendered. Copenhagen, the capital and the country's commercial and administrative centre, had been devastated by the bombardment of 1807, and Norway had been lost at the Treaty of Kiel in 1814. Trade had been seriously affected by the blockade of England, and the widespread overseas connections that formerly had played so large a part in the economic life of Denmark could not be resumed.

Copenhagen's role as an international financial and trading centre was taken over by Hamburg, whence even a considerable part of Danish home trade was controlled. Inflation further contributed to the crisis. The state was forced to make a formal declaration of bankruptcy, and not until 1818, when an independent national bank with sole rights to issue banknotes was established, was economic stability possible. It was 20 years, however, before the coinage rose to parity with the silver standard, and banknotes were first redeemable only in 1845.

The already considerable economic crisis was worsened by low grain prices. The loss of Norway and the high import duties on grain that Great Britain imposed at this time deprived Denmark of its surest markets for grain export. The agricultural crisis resulted in the compulsory auctioning of many estates and farms, forcing the agrarian reform to a complete standstill.

From 1830 economic life took a turn for the better with, among other things, more stable prices for agricultural products, increased trade, and the first signs of industrialization.

The liberal movement. Denmark's government under Frederick VI (1808–39) could be described as a patriarchal autocracy. In the Privy Council, which was regularly convened after 1814, Poul Christian Stemann became the leading figure and was responsible for the government's strongly conservative policies until 1848. His close colleague, Anders Sandøe Ørsted, pleaded for a somewhat more liberal policy, at least on economic questions. After the July Revolution in France, a demand was made in Denmark for a liberal constitution. The government was forced to make concessions, and in 1834 four provincial consultative diets (or assemblies) were created, two in the kingdom itself, one in Schleswig (Slesvig), and one in Holstein. These were not representative bodies, being composed of wealthy landholders, and their function was only advisory. As the liberal movement grew in strength, especially in the academic world and among the middle classes, the liberal press, whose leading journal was *Fædrelander* ("The Fatherland," established in 1834), subjected the monarchy and its conservative administration to severe criticism. When the popular Frederick VI died in 1839, the liberals had great hopes of his successor Christian VIII, who, during his youth as regent in Norway, had appeared as the spokesman for a liberal government. Over the years, however, he had become much more conservative and as king of Denmark did not consider the time ripe to curb the absolute monarchy. He confined himself, therefore, to modernizing the administration, especially, between 1837 and 1841, through a program of establishing local government and granting some independence to parishes and counties.

Creation of provincial diets

Parallel with the liberal movement ran the farmers' movement. This started as a religious movement but soon became dominated by social and political ideas, with agitators such as Jens Andersen Hansen leading the way. When the government intervened, the liberals and farmers joined forces against the common adversary. In 1846 the farmers' case received support when a group of liberal reformers led by Anton Frederik Tscherning founded the Society of the Friends of the Peasant (Bondevennernes Selskab), which developed into the Liberal Party (Venstre).

The 1849 constitution. After the death of Christian VIII in January 1848 and under the influence of the February



Language distribution in Schleswig-Holstein, 1849.
From W. Carr, *Schleswig-Holstein, 1815–1848*, Manchester University Press

revolution in Paris and the March revolution in Germany, the new king, Frederick VII (1848–63), installed the March Cabinet, in which the National Liberal leaders Orla Lehmann and Ditlev Gothard Monrad were given seats. After a constituent assembly had been summoned, the absolute monarchy was abolished and was replaced by the constitution of June 5, 1849. Together with the king and his ministers there was now also a Parliament with two chambers, the Folketing and the Landsting, both elected by popular vote but with a property-owning qualification as a prerequisite for a seat in the Landsting. Parliament, together with the king and government, shared legislative power while the courts independently exercised judicial power. The constitution also secured the freedom of the press, religious freedom, and the right to hold meetings and form associations.

The National Liberals and the Schleswig-Holstein question. Nationalism was, together with liberalism, the most important movement in the 19th century. In Denmark, national feelings were inflamed by the conflict with Germany on the Schleswig-Holstein question. After the loss of Norway, the Danish monarchy consisted of three main parts: the kingdom of Denmark and the duchies of Schleswig and Holstein, the last of which was a member of the German Confederation. Whereas Holstein was German, Schleswig was linguistically and culturally divided between a Danish and a German population. When the liberal German-speaking population in Schleswig opposed autocratic rule and demanded a free constitution and affiliation to Holstein and the German Confederation, a Danish National Liberal movement emerged and demanded that Schleswig be incorporated in Denmark (the Eider Policy, named for the Eider River, which formed the southern boundary of Schleswig). When the National Liberal government officially adopted this policy in 1848, the Schleswig-Holsteiners resorted to arms. The rebellion received military aid from Prussia, and the Danish army could not suppress it. The war, which lasted three years, ended in the agreements of 1851 and 1852 in which Denmark pledged to take no measures to tie Schleswig any closer to itself than to Holstein. The Eider Policy was thus abandoned, and the June constitution of 1849 applied only to Denmark.

The Eider Policy

The National Liberal government was succeeded in 1852 by a Conservative (Højre) government under Christian Albrecht Bluhme. Under the influence of the Pan-Scandinavian movement and the German Confederation's constant interference in constitutional matters in Schleswig and Holstein, the Eider Policy again won ground, and the Conservative government was replaced in 1857 by a moderate National Liberal government led by Carl Christian Hall. In 1863, in the belief that Prussia was preoccupied with the Polish rebellion against Russia and in expectation of support from Sweden, the government separated Holstein from the rest of the state and conferred a joint constitution on the kingdom and Schleswig. This "November constitution" meant that Schleswig was annexed to Denmark, in contravention of the agreements of 1851 and 1852.

Prussia, however, under the leadership of Otto von Bismarck, reacted immediately, and in February 1864 war broke out between Denmark on the one side and Prussia and Austria on the other. After the Danish defeat at Dybbøl and the consequent occupation of the whole of Jutland, Denmark was forced by the Treaty of Vienna to surrender all of Schleswig and Holstein to Prussia and Austria.

The conservative regime. *Realignment of political factions.* The defeat in 1864 led to the fall of the National Liberal government. Under Christian IX (1863–1906) a Conservative government was appointed, and in 1866 a new constitution followed that introduced electoral rules giving the Landsting a distinct conservative leaning, with great landowners and civil servants as the dominating elements. The National Liberal Party was swallowed up by the Conservative Party. As a counterweight, the various groups that represented the farmers combined together in 1870 to form the United Left (Forenede Venstre), which in 1872 secured a majority in the Folketing. The Left demanded that the 1849 June constitution be reintroduced together with a number of other reforms. With Jacob Brønnum Scavenius Estrup, a great landowner, as prime minister (1875–94), however, a strictly conservative policy was pursued. Despite the opposing parliamentary majority, the government forced its policy through by means of provisory law and with support from the king. The result was that all reforms came to a standstill. Not until the "compromise" of 1894 was the crisis solved, at which time Estrup himself left the government. The demand for parliamentary democracy was not granted, however, until the 1901 election, when the Left Reform Party (Venstre-reformparti), an offshoot of the Left, came to power, and what has become known in Denmark as the "Change of System" was introduced.

The United Left

Social and economic change. The progress of the Left and the formation of the Social Democratic Party in the 1870s must be viewed against the background of the great economic and social changes. During the 1850s and 1860s a network of railroads was created, industrialization began, agrarian reforms were introduced, and a number of technical improvements in grain production were effected. In the years between 1870 and 1901 the urban population increased from 25 percent to 44 percent of the total population. The rapid development of harbours, ships, and foreign trade meant that the shortage of raw materials, such as iron and coal, did not hinder the development of industry to any essential degree. There was a steady stream of foreign capital to Denmark. Trade unions and employers' federations were established at this time and spread nationwide by 1899. The fall in the world market price of grain after 1875 resulted in an increased production of butter and bacon. Britain became even more the main market for agricultural products. Even the smallest farm arranged its production with exports in mind, while at the same time certain foodstuffs were imported. Standardization of butter and bacon and their export was arranged by the farmers on a cooperative basis and with a view to the existing struggle with townspeople and the great landowners. The cooperative movement won ground in rural areas. Culturally, the farmers gathered around the folk high schools, educational institutions with a mainly liberal arts

program, inspired by the ideas of N.F.S. Grundtvig, the writer, educationist, and theologian.

THE 20TH CENTURY

The Left Reform government that came to power under the "Change of System" in 1901 went swiftly to work on a number of reforms. Parliamentary supremacy, requiring the king to appoint a government having the support of Parliament, began in that year. A free-trade law that corresponded to the agricultural export interests was passed; in conformity with the ideas of Grundtvig, the state church was transformed into a folk church with parochial church councils; the educational system was democratized; and changes in the system of taxation were effected so that income and not land was the main criterion for taxation. After the victory over the Conservatives, it became apparent that it was impossible for the Left Reformers, led by Jens Christian Christensen, to remain united. In 1905, therefore, the radical faction broke away to become the Radical Liberal Party (Radikale Venstre), the most important members of which were Peter Munch and Ove Rode. Between 1913 and 1920 the Radicals, supported by the Social Democrats, were in power. In 1915 the constitution was revised, and the privileged franchise to the Landsting was revoked, although the electoral qualifying age of 35 was retained. At the same time, the franchise to both the Folketing and the Landsting was extended to women, servants, and farm hands. The right-wing majority in the Landsting agreed to the constitutional reform on condition that the single-member constituency be replaced by proportional representation. There followed a number of reforms, including a judicial reform introducing trial by jury and a land reform bill that aimed to redistribute land from the large estates to increase the size of smallholders' farms.

Formation of the Radical Party

Foreign policy and World War I. After the Franco-German War (1870–71), Danish foreign policy was developed along neutral lines. There were strong differences of opinion between the Conservatives and the Left on the way in which this should be carried out. The Conservatives demanded a strong defense policy, and J.B.S. Estrup carried through the fortification of Copenhagen. Within the Left itself there was disagreement on the lines the neutrality should take. The most radical viewpoint was held by Viggo Hørup, who wanted complete disarmament. After the split within the Left in 1905, the Radical Liberal Party continued Hørup's ideas. In the years before 1914, it became increasingly important to define Germany's intended attitude to Denmark in the event of a European conflict. The Germans were well aware that the Schleswig affair had left a good many Danes with a loathing for everything German, while the constant friction between the Danish minority and the German administration in Schleswig added fuel to the flames. Danish governments after 1901 made persistent efforts to assure Germany of Denmark's benevolent neutrality, but the disagreement over this policy's implementation remained unreconciled. At the outbreak of war, Germany forced Denmark to lay mines in the Great Belt in August 1914, and, as the British fleet made no serious attempts to break through, neutrality was maintained.

World War I gave Denmark, together with a number of other neutral countries, an extremely good export market to the belligerent countries but an inevitable shortage of supplies. With a widespread overseas trade, the country's economic life was exceedingly vulnerable and became especially so after Germany opted for unrestricted submarine warfare beginning in 1917. Denmark's exports to Great Britain were, to a certain extent, reoriented to Germany. There was a shortage of raw materials in both agriculture and industry, and the government rationed a number of consumer goods and controlled the country's economic life to a certain extent.

The interwar period. By the Treaty of Versailles it was decided that part of Schleswig should revert to Denmark in accordance with the principle of self-determination. The boundary was determined by a plebiscite in 1920. The discontent that arose as a consequence of the drawing of the boundary, coupled with labour unrest and dissatisfaction with remaining wartime restrictions, led to the fall of the government in the same year. It was succeeded by a Left government supported by the Conservatives. From 1924 to 1926 the Social Democrats (Socialdemokratiet), under the leadership of Thorvald Stauning and supported by the Radicals, were in power. The years 1926 to 1929 saw the Left in power again, supported by the Conservatives.

The recurring problem for the governments of the 1920s was the critical economic conditions that followed World War I. In 1922 the country's largest private bank, Landmandsbanken, failed. At times unemployment reached a high level. The Social Democrats scored a great victory at the polls in 1929, and a coalition government was formed with the Radical Party under the leadership of the Social Democrat Stauning, with Peter Munch as foreign minister.

Monetary problems

Economic crisis. The Great Depression of the early 1930s led to unemployment, which in 1933 affected 40 percent of the organized industrial workers. When Great Britain went off the gold standard in 1931, Denmark followed suit. The greatest blow to the Danish economy, however, was the system of preferential Commonwealth tariffs established in 1932. To cope with the crisis, the government subjected foreign trade to stringent control by the establishment of a "currency centre" and won the support of the Left in the "Kanslergade compromise," by which it was agreed to devalue the krone and to freeze existing wage agreements by law. In addition, the Left supported social reforms that included old-age pensions and health, unemployment, and accident insurance. A number of measures were also adopted in support of agriculture. The general election of 1935 returned the Social Democrats again, and after the elections to the Landsting in 1936 the government coalition of Social Democrats and Radicals held the majority in both the Folketing and the Landsting. Trade improved, and during the late 1930s industry again began to expand.

Foreign policy. Denmark joined the League of Nations in 1920 and worked for a peaceful solution to international problems during the interwar period. When Adolf Hitler came to power and Germany began to rearm, Denmark's position was again vulnerable. Although Germany had never recognized the alteration in its boundary as laid down by the Treaty of Versailles, Hitler did not raise the matter. Under Foreign Minister Munch's leadership, Denmark tried in vain during the 1930s to obtain recognition of the boundary, and at the same time it avoided all measures that might possibly offend its powerful neighbour. When in 1939 Hitler offered nonaggression pacts to those countries that might feel threatened by Germany's expansionist policy, Denmark, in contrast to the other Scandinavian countries, accepted the offer.

Denmark during World War II. On the outbreak of war in 1939 Denmark, in common with the other Nordic countries, issued a declaration of neutrality. On April 9, 1940, German troops crossed the border, and after token resistance the Danish government submitted to a military occupation of the country. Formally, however, Denmark remained a sovereign state until Aug. 29, 1943, and in this its position differed from the other occupied countries of Europe. A coalition government was formed by the major parties, with Thorvald Stauning as leader, and in July 1940 Erik Scavenius became foreign minister. When Germany attacked the Soviet Union, the Danish government was forced to allow the formation of a Danish volunteer corps and at the same time to forbid all communist activity in the country. In November 1941, Denmark signed the Anti-Comintern Pact.

Stauning died suddenly in May 1942 and was succeeded by Vilhelm Buhl, who, however, was forced to resign in November of the same year under pressure from the Germans. He was succeeded by Scavenius. The elections of 1943 proved to be a great national demonstration that the people were united in support of the four old democratic parties and the fight against Nazism. At the same time, the resistance movement was organized, and Germany's military defeats paved the way for demands for an open breach with the powers of occupation. Dissatisfaction caused by consumer shortages and inflation, combined with the growing opposition to German occupation, led

Danish resistance

to a series of strikes in the summer of 1943 that in August culminated in actions aimed directly at the Germans. When the Danish government refused to introduce the death penalty for sabotage, to allow the persecution of Jews, or to use force against the strikers, the Germans declared a state of emergency. The Danish government ceased to function and the German *Reichskommissioner* assumed political control. The Danish army and navy were disbanded, but not before many of the ships were scuttled by their own crews.

In September 1943 the Danish Freedom Council was formed, and under its leadership the resistance movement was organized, mainly in the form of illegal newspapers, a comprehensive intelligence service, and numerous acts of sabotage. During the last year of the war, closer cooperation began between the Freedom Council and leading politicians. When the Germans surrendered, on May 5, 1945, a government was formed consisting half of representatives of the Freedom Council and half of politicians from the old political parties. After elections in the autumn of 1945, a Left government came to power, led by Knud Kristensen.

The postwar period. The first task after the liberation was to initiate legal proceedings against German collaborators. By a retroactive law these persons were brought to trial and sentenced to death or given long prison sentences. Another consequence of the war was that the Schleswig question arose once more. The Nazi dictatorship and the great numbers of refugees fleeing from eastern Germany to South Schleswig caused a reaction that won the Danish faction strong support among the local population. In Denmark opinion was divided, but when the United Kingdom in the autumn of 1946 made inquiries about Denmark's opinion on the boundary, all the parties agreed in the October Note of 1946 to reject any alteration of the 1920 boundary. After the elections of 1947, when Kristensen's government was replaced by a Social Democratic minority government led by Hans Hedtoft, all plans to pursue an active policy concerning the boundary question were abandoned.

Defense policy. Denmark became a member of NATO in April 1949. Denmark's military defense later was considerably strengthened by statutes in 1950 and 1951 and was further complemented by armaments from the United States. Denmark nevertheless rejected a request by the United States to establish air bases on Danish territory. With West Germany's admission to NATO, Denmark succeeded in obtaining guarantees—confirmed in the Bonn Protocol of 1955—for the rights of the Danish minority in South Schleswig. (Jö.We.)

Political developments. The constitution was substantially revised in 1953: female succession to the throne was introduced, and the national legislature was reduced to one chamber, the Folketing, whose membership was increased to 179—including two seats for Greenland and two for the Faroe Islands—based on proportional representation. The wide spectrum of political parties made it almost impossible for any one party to secure a majority. As a result governments tended to be either minority governments or coalitions of two, three, or even four parties.

Four dominant political parties

The political scene was dominated by a core of four "old" parties: the Conservatives (Konservative Folkeparti), the Left (after 1964, the Liberal Party), the Radical Liberals, and the Social Democrats. A small and changing number of other parties, such as the Communists and the Justice Party (Retsforbundet; a single-tax party based on the ideas of Henry George), complicated the political and parliamentary situation.

From 1953 to 1968 the Social Democrats were in power, either alone or in coalition with the Radicals and, for a short period, the Justice Party, and always with a Social Democrat as prime minister. The major results were new tax laws: a general value-added consumer tax and income taxes deducted as earned were introduced, enabling the government to stimulate or restrain demand by lowering or raising the level of taxation.

In the 1968 election the majority shifted to the right. The Radical Liberals' leader, Hilmar Baunsgaard, deserted the Social Democrats and headed a coalition with the Conservatives and the Liberals until 1971, when Jens Otto Krag again formed a Social Democratic government.

On Jan. 14, 1972, King Frederick IX died, and his eldest daughter was proclaimed queen as Margrethe II.

The day after the referendum on Oct. 2, 1972, in which 63 percent of the voters approved Danish membership in the European Economic Community (EEC), Krag unexpectedly resigned, leaving the post of prime minister to Anker Jørgensen, who had to call an election in November 1973. An electoral landslide resulted in heavy losses for the four "old" parties and the emergence of three new parties: the Centre Democrats (Centrum-Demokraterne), the Christian People's Party (Kristeligt Folkeparti), and the Progress Party (Fremskridtspartiet), an antitax party. A weak minority government under Poul Hartling of the Liberal Party tried to solve the country's growing economic problems, but his austerity program resulted in protests from trade unions and the opposition. He appealed to the country in January 1975, but Jørgensen again came to power (from 1978 in coalition with the Liberals), rejecting support from the left-wing Socialist People's Party (Socialistisk Folkeparti), which opposed Danish membership in NATO.

The end of the 1970s brought a deteriorating economic situation and the political system's inability to reach a consensus on measures to solve the problems. Increased indirect taxes to reduce the foreign debt and the deficit on the balance of payments met with strong opposition from the trade unions; in 1979 Jørgensen was again forced to resign after the two parties had failed to agree on how to implement a price and income policy. After the election in October, however, he formed a Social Democratic minority government, which introduced what was called the most stringent wage-and-price-freeze program since World War II.

After a new general election in December 1981, the voting age having been reduced from 20 to 18 following a referendum, Jørgensen again lost seats in the Folketing, but he continued as leader of a weak minority government that faced many problems, especially high unemployment, which had risen to about 10 percent. He was once more forced to resign—this time, however, without an election—in September 1982. The leader of the Conservative Party, Poul Schlüter, formed a minority government with three other centre-right parties: the Liberals, the Centre Democrats, and the Christian People's Party. Together, they had only 66 seats in the Folketing.

The start of Schlüter's decade

Schlüter, the first Conservative prime minister since 1901, introduced a counterinflationary and economic recovery program that yielded results in 1985–86, but the country's foreign debt and balance-of-payments deficit continued to cause serious concern during the 1980s. Schlüter was consequently forced to call several general elections (1984, 1987, 1988), carry out government reshuffles (1986, 1987, 1988, 1989), and threaten to call elections or to resign. He survived 23 no-confidence votes concerning foreign and defense policy, brought by the Social Democrats in tactical attempts to force him from office.

When Schlüter reshuffled the government in 1988, he incorporated the Radical Liberals and excluded the Christian People's Party and the Centre Democrats. The coalition government came under greater pressure from the left-wing Socialist People's Party and the right-wing Progress Party, both of which gained seats in the Folketing at the end of the 1980s; the Progress Party advocated substantial cuts in the public sector and a more restrictive policy toward the dramatically increased number of refugees. It was a scandal over Tamil refugees that forced Schlüter's resignation in 1993 and brought a coalition government under the leadership of Social Democrat Poul Nyrup Rasmussen to power. The early 1990s also brought a gradual recovery in the Danish economy, despite the general European recession.

Economic issues. The domestic scene since the mid-20th century has been dominated by intermittently severe economic problems. From the 1950s onward the frequently negative balance of payments, the labour market, and the country's trade policy were troubling economic and political issues.

Imbalance of trade

Although during the early 1950s the Danish economy suffered a large deficit in the trade balance, the situation improved during the later 1950s as the result of lower import prices for raw materials, a considerable increase in industrial production, and the stabilization of prices of agricultural export products. The period from 1957 to 1965 saw rapidly rising prosperity. Within the framework of the Organisation for European Economic Co-operation (OEEC), Denmark, during the 1950s, abolished most of the regulations that had restricted its foreign trade, and it was one of the founding members of the European Free Trade Association (EFTA) in 1959. In 1972 Denmark was offered, and accepted, EEC membership, which became effective on Jan. 1, 1973.

During the 1960s, however, the balance-of-payments deficit became larger, and the government was forced to intervene in an attempt to control rising consumption. This was done by adding a purchase tax in 1962, by compulsory savings, by intervention in labour conflicts, and by the regulation of wages and prices. Economic problems worsened in the 1970s. The governments attempted to impose stringent measures such as harsh savings programs, but strong opposition to various plans led to the dissolution of the Folketing on several occasions. After 1973 rising oil prices and the international recession affected the Danish economy badly and led to a dramatic increase in unemployment.

In the 1980s the government was forced to impose several austerity measures which resulted in a record high level of taxation. The measures yielded results: lower inflation, recovery in business confidence and investments, growth of employment in the private sector, and increasing economic activity. It proved difficult, however, to eliminate the budget deficit, and in 1986 the government was forced to impose new austerity measures. Balance-of-payments deficits and persistent unemployment, however, continued to plague Denmark throughout the 1980s. (M.I.A.L./Ed.)

During the 1990s, while the economy improved and unemployment dropped, Danes struggled with three key political and economic issues. Political controversy surrounded the status of immigrants and refugees in Denmark. A violation of refugees' rights caused a conservative government to fall in 1993, right-wing parties adopted anti-immigration platforms, and rioting followed the expulsion in 1999 from Denmark of a Danish-born man of Turkish descent. While most Danes supported maintaining the country's strong social welfare programs, some Danes sought to decrease the programs' high cost in taxes, while others opposed any cuts in benefits. Danes also were divided during the 1990s over closer economic ties with the European Community. In 1992 Danish voters rejected the Maastricht Treaty, which provided the framework for an expanded European Union (EU) that would subsume the EC. A second referendum in 1993 approved Danish membership in the EU, but only after Denmark had negotiated exemptions from certain provisions of the treaty, which many Danes thought might erode Danish social benefits or environmental protections. In a 2000 referendum, Danish voters rejected the single European currency, the euro. Despite these controversies, Denmark's economy prospered at the turn of the 21st century, and the strength of its information and environmental technologies promised a bright future for the country. (Ed.)

For later developments in the history of Denmark, see the BRITANNICA BOOK OF THE YEAR.

For coverage of related topics in the *Macropædia* and *Micropædia*, see the *Propædia*, sections 923, 961, 963, and 972, and the *Index*.

BIBLIOGRAPHY

Physical and human geography: KENNETH E. MILLER (comp.), *Denmark* (1987), contains an annotated bibliography of various 19th- and 20th-century publications. ED THOMASSON, *Danish Quality Living: The Good Life Handbook* (1985), provides a casual introduction to how Danes sometimes describe themselves to foreigners. JUDITH FRIEDMAN HANSEN, *We Are a Little Land: Cultural Assumptions in Danish Everyday Life* (1980), describes the social and cultural values that characterize the Danish lifestyle as a distinctive variant of modern Euro-American civilization. ROBERT T. ANDERSON and BARBARA GALLATIN ANDERSON, *The Vanishing Village: A Danish Maritime Community* (1964), is an easy-to-read study of Danish life as it changed from that of a small inner-focused community to that of a mid-20th-century suburb of Copenhagen. CLEMENS PEDERSEN (ed.), *The Danish Co-operative Movement*, trans. from Danish (1977), offers an authoritative history of how Danish cooperatives first became influential in shaping the modernization of agriculture in Denmark and how they now function. THOMAS RØRDAM, *The Danish Folk High Schools*, 2nd rev. ed., trans. from Danish (1980), describes historically the movement initiated by N.S.F. Grundtvig that culminated in the folk high school movement as a means of putting education to the service of defining national goals of equality and self-respect for a peasant ancestry. ERIC S. EINHORN and JOHN LOGUE, *Modern Welfare States: Politics and Policies in Social Democratic Scandinavia* (1989), extensively describes and analyzes the expansion of the public sector in developing and managing the social welfare system that characterizes Denmark, Norway, and Sweden. ERIK ALLARDT *et al., Nordic Democracy* (1981), is a well-documented, densely informative description of political institutions in Scandinavia. STANLEY V. ANDERSON, *The Nordic Council: A Study of Scandinavian Regionalism* (1967), a rather technical account from the perspective of political science and international law, studies how Danish communal values find expression through international cooperation with other Scandinavian nations.

History: General works include STEWART OAKLEY, *A Short History of Denmark* (also published as *The Story of Denmark*, 1972), a readable work; W. GLYN JONES, *Denmark: A Modern History*, rev. ed. (1986), a well-written survey; PALLE LAURING, *A History of Denmark*, 7th ed. (1986); and BENT RYING, *Danish in the South and the North*, vol. 2, *Denmark: History*, trans. from Danish (1988), which deals with the development from the Stone Age to present times, with excellent pictures. For more advanced studies, consult OLAF OLSEN (ed.), *Gyldendal og Politikens Danmarkshistorie* (1988–), with 7 vol. published by 1989; and the series "Dansk socialhistorie," 7 vol. (1979–82), on social history from the Stone Age to 1978—vol. 1 has appeared in an English trans. as *The Prehistory of Denmark*, by JØRGEN JENSEN (1982). Danish prehistory and archaeology are examined by PALLE LAURING, *Land of the Tollund Man* (1957; originally published in Danish, 1954), covering the first settlers of hunting nomads to the Vikings; P.V. GLOB, *Denmark: An Archaeological History from the Stone Age to the Vikings* (also published as *Danish Prehistoric Monuments*, 1971; originally published in Danish, 1942), a scholarly survey, *The Mound People: Danish Bronze-Age Man Preserved* (1974, reissued 1983; originally published in Danish, 1970), a thoroughly illustrated technical monograph, and *The Bog People: Iron Age Man Preserved* (1969, reissued 1988; originally published in Danish, 1965); and FLSE ROESDAHL, *Viking Age Denmark* (1982; originally published in Danish, 1980), an extensive description of Viking activities. RUTH MAZO KARRAS, *Slavery and Society in Medieval Scandinavia* (1988), draws on a wide variety of primary sources and archaeological data about the social, legal, and economic aspects of slavery. SVEND ELLEHØJ (ed.), *Christian IV's verden* (1988), correlates the findings and views of modern scholarship on the king and his times. SVEND AAGE HANSEN, *Økonomisk vækst i Danmark*, 2 vol. (1972–74), gives a broad view of the economic growth in the period 1720–1970, with statistics. FRIDLEV SKRUBBELTRANG, *Den danske Landbovæmfund 1500–1800* (1978), concerns agriculture. JØRGEN HÆSTRUP, *Secret Alliance: A Study of the Danish Resistance Movement, 1940–1945*, 3 vol. (1976–77; originally published in Danish, 1954), analyzes the movement in detail, based on "illegal" documents and personal accounts by leading members of the Resistance. HARRY HAUE, JØRGEN OLSEN, and JØRN AARUP-KRISTENSEN, *Det ny Danmark 1890–1985: Udviklingslinjer og tendens*, 3rd ed. (1985), deals with modern history. For current research, three journals are useful: *The Scandinavian Economic History Review* (3/yr.); *Scandinavian Journal of History* (quarterly); and *Scandinavian Political Studies* (quarterly).

(R.T.A./S.V.A./M.I.A.L.)

The Great Depression

A worldwide economic downturn that began in 1929 and lasted until about 1939, the Great Depression was unique for its duration and for the number of countries it affected. It was the longest and most severe depression ever experienced by the industrialized Western world. Responses to the economic hardships of the era sparked fundamental changes in political and economic institutions, macroeconomic policy, and economic theory. Although it originated in the United States, the Great Depression caused drastic declines in output, severe unemployment, and acute deflation in almost every country of the world. Its social and cultural effects were no less staggering, especially in the United States, where the Great Depression represented the harshest adversity faced by Americans since the Civil War.

Economic history 244
Timing and severity 244
Causes of the decline 245
Stock market crash
Banking panics and monetary contraction
The gold standard
International lending and trade
Sources of recovery 245A
Economic impact 245B
Culture and society in the Great Depression 245C
Global concerns 245C
Political movements and social change 245C
New forms of cultural expression 245D
The documentary impulse
Federal arts programs
Theatre
Fiction
Popular culture
Portrayals of hope 245G
Bibliography 245H

Economic history

The timing and severity of the Great Depression varied substantially across countries. The Depression was particularly long and severe in the United States and Europe; it was milder in Japan and much of Latin America. Perhaps not surprisingly, the worst depression ever experienced by the world economy stemmed from a multitude of causes. Declines in consumer demand, financial panics, and misguided government policies caused economic output to fall in the United States, while the gold standard, which linked nearly all the countries of the world in a network of fixed currency exchange rates, played a key role in transmitting the American downturn to other countries. The recovery from the Great Depression was spurred largely by the abandonment of the gold standard and the ensuing monetary expansion. The economic impact of the Great Depression was enormous, including both extreme human suffering and profound changes in economic policy.

Multiple causes

TIMING AND SEVERITY

The Great Depression began in the United States as an ordinary recession in the summer of 1929. The downturn became markedly worse, however, in late 1929 and continued until early 1933. Real output and prices fell precipitously. Between the peak and the trough of the downturn, industrial production in the United States declined 47 percent and real gross domestic product (GDP) fell 30 percent. The wholesale price index declined 33 percent (such declines in the price level are referred to as deflation). Although there is some debate about the reliability of the statistics, it is widely agreed that the unemployment rate exceeded 20 percent at its highest point. The severity of the Great Depression in the United States becomes especially clear when it is compared with America's next worst recession of the 20th century, that of 1981–82, when the country's real GDP declined just 2 percent and the unemployment rate peaked at less than 10 percent. Moreover, during the 1981–82 recession prices continued to rise, although the rate of price increase slowed substantially (a phenomenon known as disinflation).

Table 1: Dates of the Great Depression in Various Countries (in quarters)

country	depression began	recovery began
United States	1929:3	1933:2
United Kingdom	1930:1	1932:4
Germany	1928:1	1932:3
France	1930:2	1932:3
Italy	1929:3	1933:1
Japan	1930:1	1932:3
Canada	1929:2	1933:2
Belgium	1929:3	1932:4
The Netherlands	1929:4	1933:2
Sweden	1930:2	1932:3
Switzerland	1929:4	1933:1
Denmark	1930:4	1933:2
Poland	1929:1	1933:2
Czechoslovakia	1929:4	1933:2
Argentina	1929:2	1932:1
Brazil	1928:3	1931:4
India	1929:4	1931:4
South Africa	1930:1	1933:1

The Depression affected virtually every country of the world. However, the dates and magnitude of the downturn varied substantially across countries. Table 1 shows the dates of the downturn and the upturn in economic activity in a number of countries. Table 2 shows the peak-to-trough percentage decline in annual industrial production for countries for which such data are available. Great Britain struggled with low growth and recession during most of the second half of the 1920s. Britain did not slip into severe depression, however, until early 1930, and its peak-to-trough decline in industrial production was roughly one-third that of the United States. France also experienced a relatively short downturn in the early 1930s. The French recovery in 1932 and 1933, however, was short-lived. French industrial production and prices both fell substantially between 1933 and 1936. Germany's economy slipped into a downturn early in 1928 and then stabilized before turning down again in the third quarter of 1929. The decline in German industrial production was roughly equal to that in the United States. A number of countries in Latin America fell into depression in late 1928 and early 1929, slightly before the U.S. decline in output. While some less-developed countries experienced severe depres-

Global downturns in prices and production

Table 2: Peak-to-Trough Decline in Industrial Production in Various Countries (annual data)

country	percent decline
United States	46.8
United Kingdom	16.2
Germany	41.8
France	31.3
Italy	33.0
Japan	8.5
Canada	42.4
Belgium	30.6
The Netherlands	37.4
Sweden	10.3
Denmark	16.5
Poland	46.6
Czechoslovakia	40.4
Argentina	17.0
Brazil	7.0







(Upper left) Nazi Storm Troopers marching through the streets of Nurnberg, Germany, September 1933.

(Upper right) Dust storm in Baca county, Colorado, 1936.

(Centre left) Flood victims waiting for food and clothing from the Red Cross; photo by Margaret Bourke-White, 1937.

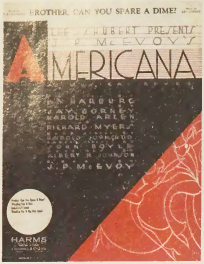
(Below left) Shantytowns such as this one in Seattle, Washington, were known as "Hoovervilles."

(Below right) New York City's financial district on Black Thursday, October 24, 1929.





(Top left) akg-images, London; (top right) Library of Congress, Washington, D.C.; (centre) Margaret Bourke-White—Time Life Pictures/Getty Images; (bottom left, bottom right) © Bettmann/Corbis



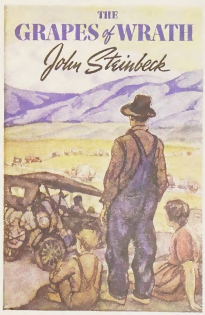
Sheet music for "Brother, Can You Spare A Dime?" by E.Y. Harburg and Jay Gorney, 1932.



Americans in Madrid during the Spanish Civil War: (left to right) writer John Dos Passos, filmmaker Joris Evans, bullfighter Sidney Franklin, and writer Ernest Hemingway.



Ginger Rogers and Fred Astaire dancing in the film *Swing Time* (1936).



Cover of John Steinbeck's novel *The Grapes of Wrath* (1939); artwork by Elmer Hader.



"Years of Dust," colour lithograph by Ben Shahn, 1936.

(Top left) The Lester S. Levy Collection of Sheet Music, Special Collections, The Milton S. Eisenhower Library of The Johns Hopkins University; (top right) © Bettmann/Corbis; (centre left) RKO Radio Pictures/The Kobal Collection; (bottom right) Art © Estate of Ben Shahn/Licensed by VAGA, New York, New York; photograph, Library of Congress, Washington, D.C.; (bottom left) Viking Press/Penguin Group (USA) Inc./Between the Covers Rare Books, Inc., Merchantville, New Jersey

sions, others, such as Argentina and Brazil, experienced comparatively mild downturns. Japan also experienced a mild depression, which began relatively late and ended relatively early.

The general price deflation evident in the United States was also present in other countries. Virtually every industrialized country endured declines in wholesale prices of 30 percent or more between 1929 and 1933. Because of the greater flexibility of the Japanese price structure, deflation in Japan was unusually rapid in 1930 and 1931. This rapid deflation may have helped to keep the decline in Japanese production relatively mild. The prices of primary commodities traded in world markets declined even more dramatically during this period. For example, the prices of coffee, cotton, silk, and rubber were reduced by roughly half just between September 1929 and December 1930. As a result, the terms of trade declined precipitously for producers of primary commodities.

The U.S. recovery began in the spring of 1933. Output grew rapidly in the mid-1930s: real GDP rose at an average rate of 9 percent per year between 1933 and 1937. Output had fallen so deeply in the early years of the 1930s, however, that it remained substantially below its long-run trend path throughout this period. In 1937–38 the United States suffered another severe downturn, but after mid-1938 the American economy grew even more rapidly than in the mid-1930s. The country's output finally returned to its long-run trend path in 1942.

A second economic downturn

Recovery in the rest of the world varied greatly. The British economy stopped declining soon after Great Britain abandoned the gold standard in September 1931, although genuine recovery did not begin until the end of 1932. The economies of a number of Latin American countries began to strengthen in late 1931 and early 1932. Germany and Japan both began to recover in the fall of 1932. Canada and many smaller European countries started to revive at about the same time as the United States, early in 1933. On the other hand, France, which experienced severe depression later than most countries, did not firmly enter the recovery phase until 1938.

CAUSES OF THE DECLINE

The fundamental cause of the Great Depression in the United States was a decline in spending (sometimes referred to as aggregate demand), which led to a decline in production as manufacturers and merchandisers noticed an unintended rise in inventories. The sources of the contraction in spending in the United States varied over the course of the Depression, but they cumulated in a monumental decline in aggregate demand. The American decline was transmitted to the rest of the world largely through the gold standard. However, a variety of other factors also influenced the downturn in various countries.

Stock market crash. The initial decline in U.S. output in the summer of 1929 is widely believed to have stemmed from tight U.S. monetary policy aimed at limiting stock market speculation. The 1920s had been a prosperous decade but not an exceptional boom period; prices had remained nearly constant throughout the decade, and there had been mild recessions in both 1924 and 1927. The one obvious area of excess was the stock market. Stock prices had risen more than fourfold from the low in 1921 to the peak in 1929. In 1928 and 1929 the Federal Reserve had raised interest rates in hopes of slowing the rapid rise in stock prices. These higher interest rates depressed interest-sensitive spending in areas such as construction and automobile purchases, which in turn reduced production. Some scholars believe that a boom in housing construction in the mid-1920s led to an excess supply of housing and a particularly large drop in construction in 1928 and 1929.

Excessive stock speculation

By the fall of 1929, U.S. stock prices had reached levels that could not be justified by reasonable anticipations of future earnings. As a result, when a variety of minor events led to gradual price declines in October 1929, investors lost confidence and the stock market bubble burst. Panic selling began on "Black Thursday," October 24, 1929. Many stocks had been purchased on margin—that is, using loans secured by only a small fraction of the stocks' value. As a result, the price declines forced some investors to liquidate their holdings, thus exacerbating the fall in prices. Between their peak in September and their low in November, U.S. stock prices (measured by the Cowles Index) declined 33 percent. Because the decline was so dramatic, this event is often referred to as the Great Crash of 1929.

The stock market crash reduced American aggregate demand substantially. Consumer purchases of durable goods and business investment fell sharply after the crash. A likely explanation is that the financial crisis generated considerable uncertainty about future income, which in turn led consumers and firms to put off purchases of durable goods. Although the loss of wealth caused by the decline in stock prices was relatively small, the crash may also have depressed spending by making people feel poorer. As a result of the drastic decline in consumer and business spending, real output in the United States, which had been declining slowly up to this point, fell rapidly in late 1929 and throughout 1930. Thus, while the Great Crash of the stock market and the Great Depression are two quite separate events, the decline in stock prices was one factor contributing to declines in production and employment in the United States.

Banking panics and monetary contraction. The next blow to aggregate demand occurred in the fall of 1930, when the first of four waves of banking panics gripped the United States. A banking panic arises when many depositors simultaneously lose confidence in the solvency of banks and demand that their bank deposits be paid to them in cash. Banks, which typically hold only a fraction of deposits as cash reserves, must liquidate loans in order to raise the required cash. This process of hasty liquidation can cause even a previously solvent bank to fail. The United States experienced widespread banking panics in the fall of 1930, the spring of 1931, the fall of 1931, and the fall of 1932. This final wave of panics continued through the winter of 1933 and culminated with the national "bank holiday" declared by President Franklin D. Roosevelt on March 6, 1933. The bank holiday closed all banks, and they were permitted to reopen only after being deemed solvent by government inspectors. The panics took a severe toll on the American banking system. By 1933, one-fifth of the banks in existence at the start of 1930 had failed.

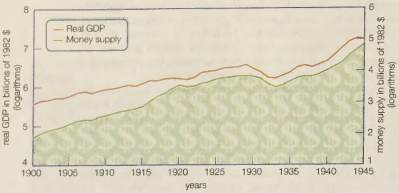
Runs on banks

By their nature, banking panics are largely irrational, inexplicable events, but some of the factors contributing to the problem can be explained. Economic historians believe that substantial increases in farm debt in the 1920s, together with U.S. policies that had encouraged small, undiversified banks, created an environment in which such panics could ignite and spread. The heavy farm debt stemmed in part from the high prices of agricultural goods during World War I, which had spurred extensive borrowing by American farmers wishing to increase production by investing in land and machinery. The decline in farm commodity prices following the war made it difficult for farmers to keep up with their loan payments.

The Federal Reserve did little to try to stem the banking panics. Economists Milton Friedman and Anna J. Schwartz, in the classic study *A Monetary History of the United States, 1867–1960* (1963), argued that the death in 1928 of Benjamin Strong, who had been the governor of the Federal Reserve Bank of New York since 1914, was a significant cause of this inaction. Strong had been a forceful leader who understood the ability of the central bank to limit panics. His death left a power vacuum at the Federal Reserve and allowed leaders with less sensible views to block effective intervention. The panics caused a dramatic rise in the amount of currency people wished to hold relative to their bank deposits. This rise in the currency-to-deposit ratio was a key reason why the money supply in the United States declined 31 percent between 1929 and 1933. In addition to allowing the panics to reduce the U.S. money supply, the Federal Reserve also deliberately contracted the money supply and raised interest rates in September 1931, when Britain was forced off the gold standard and investors feared that the United States would devalue as well.

Decline in the money supply

Scholars believe that such declines in the money supply caused by Federal Reserve decisions had a severely con-



Money and output in the United States, 1900–1945.

tractionary effect on output. A simple picture provides perhaps the clearest evidence of the key role monetary collapse played in the Great Depression in the United States. The figure shows the money supply and real output over the period 1900 to 1945. In ordinary times, such as the 1920s, both the money supply and output tend to grow steadily. But in the early 1930s both plummeted. The decline in the money supply depressed spending in a number of ways. Perhaps most important, because of actual price declines and the rapid decline in the money supply, consumers and businesspeople came to expect deflation; that is, they expected wages and prices to be lower in the future. As a result, even though nominal interest rates were very low, people did not want to borrow because they feared that future wages and profits would be inadequate to cover their loan payments. This hesitancy, in turn, led to severe reductions in both consumer spending and business investment spending. The panics surely exacerbated the decline in spending by generating pessimism and loss of confidence. Furthermore, the failure of so many banks disrupted lending, thereby reducing the funds available to finance investment.

Expectations of continued deflation

The gold standard. Some economists believe that the Federal Reserve allowed or caused the huge declines in the American money supply partly to preserve the gold standard. Under the gold standard, each country set the value of its currency in terms of gold and took monetary actions to defend the fixed price. It is possible that, had the Federal Reserve expanded the money supply greatly in response to the banking panics, foreigners could have lost confidence in the United States' commitment to the gold standard. This could have led to large gold outflows, and the United States could have been forced to devalue. Likewise, had the Federal Reserve not tightened in the fall of 1931, it is possible that there would have been a speculative attack on the dollar and the United States would have been forced to abandon the gold standard along with Great Britain.

While there is debate about the role the gold standard played in limiting U.S. monetary policy, there is no question that it was a key factor in the transmission of America's economic decline to the rest of the world. Under the gold standard, imbalances in trade or asset flows gave rise to international gold flows. For example, in the mid-1920s intense international demand for American assets such as stocks and bonds brought large inflows of gold to the United States. Likewise, a decision by France after World War I to return to the gold standard with an undervalued franc led to trade surpluses and substantial gold inflows. (See also TRADE, BALANCE OF in the *Micropædia*.)

Britain chose to return to the gold standard after World War I at the prewar parity. Wartime inflation, however, implied that the pound was overvalued, and this overvaluation led to trade deficits and substantial gold outflows after 1925. To stem the gold outflow, the Bank of England raised interest rates substantially. High interest rates depressed British spending and led to high unemployment in Great Britain throughout the second half of the 1920s.

Once the U.S. economy began to contract severely, the tendency for gold to flow out of other countries and toward the United States intensified. This took place because deflation in the United States made American goods particularly desirable to foreigners, while low income reduced American demand for foreign products. To counteract the resulting tendency toward an American trade surplus and foreign gold outflows, central banks throughout the world raised interest rates. Maintaining the international gold standard, in essence, required a massive monetary contraction throughout the world to match the one occurring in the United States. The result was a decline in output and prices in countries throughout the world that also nearly matched the downturn in the United States.

Effects of the gold standard

Financial crises and banking panics occurred in a number of countries besides the United States. In May 1931 payment difficulties at the Creditanstalt, Austria's largest bank, set off a string of financial crises that enveloped much of Europe and were a key factor forcing Britain to abandon the gold standard. Among the countries hardest hit by bank failures and volatile financial markets were Austria, Germany, and Hungary. These widespread banking crises could have been the result of poor regulation and other local factors or of simple contagion from one country to another. In addition, the gold standard, by forcing countries to deflate along with the United States, reduced the value of banks' collateral and made them more vulnerable to runs. As in the United States, banking panics and other financial market disruptions further depressed output and prices in a number of countries.

International lending and trade. Some scholars stress the importance of other international linkages. Foreign lending to Germany and Latin America had expanded greatly in the mid-1920s, but U.S. lending abroad fell in 1928 and 1929 as a result of high interest rates and the booming stock market in the United States. This reduction in foreign lending may have led to further credit contractions and declines in output in borrower countries. In Germany, which experienced extremely rapid inflation ("hyperinflation") in the early 1920s, monetary authorities may have hesitated to undertake expansionary policy to counteract the economic slowdown because they worried it might reignite inflation. The effects of reduced foreign lending may explain why the economies of Germany, Argentina, and Brazil turned down before the Great Depression began in the United States.

The 1930 enactment of the Smoot-Hawley tariff in the United States and the worldwide rise in protectionist trade policies created other complications. The Smoot-Hawley tariff was meant to boost farm incomes by reducing foreign competition in agricultural products. But other countries followed suit, both in retaliation and in an attempt to force a correction of trade imbalances. Scholars now believe that these policies may have reduced trade somewhat but were not a significant cause of the Depression in the large industrial producers. Protectionist policies, however, may have contributed to the extreme decline in the world price of raw materials, which caused severe balance-of-payments problems for primary-commodity-producing countries in Africa, Asia, and Latin America and led to contractionary monetary and fiscal policies.

SOURCES OF RECOVERY

Given the key roles of monetary contraction and the gold standard in causing the Great Depression, it is not surprising that currency devaluations and monetary expansion were the leading sources of recovery throughout the world. There is a notable correlation between the times at which countries abandoned the gold standard (or devalued their currencies substantially) and when they experienced renewed growth in their output. For example, Britain, which was forced off the gold standard in September 1931, recovered relatively early, while the United States, which did not effectively devalue its currency until 1933, recovered substantially later. Similarly, the Latin American countries of Argentina and Brazil, which began to devalue in 1929, experienced relatively mild downturns and had largely recovered by 1935. In contrast, the "Gold Bloc" countries of Belgium and France, which were particularly wedded to the gold standard and slow to devalue, still had industrial production in 1935 well below that of 1929.

Role of monetary policy

Devaluation, however, did not increase output directly. Rather, it allowed countries to expand their money supplies without concern about gold movements and ex-

change rates. Countries that took greater advantage of this freedom saw greater recovery. The monetary expansion that began in the United States in early 1933 was particularly dramatic. The American money supply increased nearly 42 percent between 1933 and 1937. This monetary expansion stemmed largely from a substantial gold inflow to the United States, caused in part by the rising political tensions in Europe that eventually led to World War II. Monetary expansion stimulated spending by lowering interest rates and making credit more widely available. It also created expectations of inflation, rather than deflation, thereby giving potential borrowers greater confidence that their wages and profits would be sufficient to cover their loan payments if they chose to borrow. One sign that monetary expansion stimulated recovery in the United States by encouraging borrowing was that consumer and business spending on interest-sensitive items such as cars, trucks, and machinery rose well before consumer spending on services.

Fiscal policy played a relatively small role in stimulating recovery in the United States. Indeed, the Revenue Act of 1932 increased American tax rates greatly in an attempt to balance the federal budget, and by doing so it dealt another contractionary blow to the economy by further discouraging spending.

Federal programs

Franklin D. Roosevelt's New Deal, initiated in early 1933, did include a number of new federal programs aimed at generating recovery. For example, the Works Progress Administration (WPA) hired the unemployed to work on government building projects, and the Tennessee Valley Authority (TVA) constructed dams and power plants in a particularly depressed area. However, the actual increases in government spending and the government budget deficit were small relative to the size of the economy. This is especially apparent when state government budget deficits are included, because those deficits actually declined at the same time that the federal deficit rose. As a result, the new spending programs initiated by the New Deal had little direct expansionary effect on the economy. Whether they may nevertheless have had positive effects on consumer and business sentiment remains an open question.

Some New Deal programs may have actually hindered recovery. The National Industrial Recovery Act of 1933, for example, set up the National Recovery Administration (NRA), which encouraged firms in each industry to adopt a code of behaviour. These codes discouraged price competition between firms, set minimum wages in each industry, and sometimes limited production. Likewise, the Agricultural Adjustment Act of 1933 created the Agricultural Adjustment Administration (AAA), which set voluntary guidelines and gave incentive payments to farmers to restrict production in hopes of raising agricultural prices. Modern research suggests that such anticompetitive practices and wage and price guidelines led to inflation in the early recovery period in the United States and discouraged reemployment and production.

Recovery in the United States was stopped short by another distinct recession that began in May 1937 and lasted until June 1938. One source of the 1937–38 recession was a decision by the Federal Reserve to greatly increase reserve requirements. This move, which was prompted by fears that the economy might be developing speculative excess, caused the money supply to cease its rapid growth and to actually fall again. Fiscal contraction and a decrease in inventory investment due to labour unrest are also thought to have contributed to the downturn. That the United States experienced a second, very severe contraction before it had completely recovered from the enormous decline of the early 1930s is the main reason that the United States remained depressed for virtually the entire decade.

World War II played only a modest role in the recovery of the U.S. economy. Despite the recession of 1937–38, real GDP in the United States was well above its pre-Depression level by 1939, and by 1941 it had recovered to within about 10 percent of its long-run trend path. Therefore, in a fundamental sense, the United States had largely recovered before military spending accelerated noticeably. At the same time, the U.S. economy was still somewhat below trend at the start of the war, and the unemployment rate averaged just under 10 percent in 1941. The government budget deficit grew rapidly in 1941 and 1942 because of the military buildup, and the Federal Reserve responded to the threat and later the reality of war by increasing the money supply greatly over the same period. This expansionary fiscal and monetary policy, together with widespread conscription beginning in 1942, quickly returned the economy to its trend path and reduced the unemployment rate to below its pre-Depression level. So, while the war was not the main impetus for the recovery in the United States, it played a role in completing the return to full employment.

The role of fiscal expansion, and especially of military expenditure, in generating recovery varied substantially across countries. Great Britain, like the United States, did not use fiscal expansion to a noticeable extent early in its recovery. It did, however, increase military spending substantially after 1937. France raised taxes in the mid-1930s in an effort to defend the gold standard but then ran large budget deficits starting in 1936. The expansionary effect of these deficits, however, was counteracted somewhat by a legislated reduction in the French workweek from 46 to 40 hours—a change that raised costs and depressed production. Fiscal policy was used more successfully in Germany and Japan. The German budget deficit as a percent of domestic product increased little early in the recovery, but it grew substantially after 1934 as a result of spending on public works and rearmament. In Japan, government expenditures, particularly military spending, rose from 31 to 38 percent of domestic product between 1932 and 1934, resulting in substantial budget deficits. This fiscal stimulus, combined with substantial monetary expansion and an undervalued yen, returned the Japanese economy to full employment relatively quickly.

ECONOMIC IMPACT

The most devastating impact of the Great Depression was human suffering. In a short period of time, world output and standards of living dropped precipitously. As much as one-fourth of the labour force in industrialized countries was unable to find work in the early 1930s. While conditions began to improve by the mid-1930s, total recovery was not accomplished until the end of the decade.

The Great Depression and the policy response also changed the world economy in crucial ways. Most obviously, it hastened, if not caused, the end of the international gold standard. Although a system of fixed currency exchange rates was reinstated after World War II under the Bretton Woods system, the economies of the world never embraced that system with the conviction and fervour they had brought to the gold standard. By 1973, fixed exchange rates had been abandoned in favour of floating rates. (See also MONEY.)

Growth of labour unions

Both labour unions and the welfare state expanded substantially during the 1930s. In the United States, union membership more than doubled between 1930 and 1940. This trend was stimulated by both the severe unemployment of the 1930s and the passage of the National Labor Relations (Wagner) Act (1935), which encouraged collective bargaining. The United States also established unemployment compensation and old-age and survivors' insurance through the Social Security Act (1935), which was passed in response to the hardships of the 1930s. It is uncertain whether these changes would have eventually occurred in the United States without the Great Depression. Many European countries had experienced significant increases in union membership and had established government pensions before the 1930s. Both of these trends, however, accelerated in Europe during the Great Depression.

In many countries, government regulation of the economy, especially of financial markets, increased substantially in the 1930s. The United States, for example, established the Securities and Exchange Commission (1934) to regulate new stock issues and stock market trading practices. The Banking Act of 1933 (also known as the Glass-Steagall Act) established deposit insurance in the United States and prohibited banks from underwriting or dealing in securi-

ties. Deposit insurance, which did not become common worldwide until after World War II, effectively eliminated banking panics as an exacerbating factor in recessions in the United States after 1933.

The Great Depression also played a crucial role in the development of macroeconomic policies intended to temper economic downturns and upturns. The central role of reduced spending and monetary contraction in the Depression led British economist John Maynard Keynes to develop the ideas in his *General Theory of Employment, Interest, and Money* (1936). Keynes's theory suggested that increases in government spending, tax cuts, and monetary expansion could be used to counteract depressions. This insight, combined with a growing consensus that government should try to stabilize employment, has led to much more activist policy since the 1930s. Legislatures and central banks throughout the world now routinely attempt to prevent or moderate recessions. Whether such a change would have occurred without the Depression is again a largely unanswerable question. What is clear is that this change has made it unlikely that a decline in spending will ever be allowed to multiply and spread throughout the world as it did during the Great Depression of the 1930s.

Stabilization policy

(C.D.Ro.)

Culture and society in the Great Depression

No decade in the 20th century was more terrifying for people throughout the world than the 1930s. The traumas of the decade included economic disorder, the rise of totalitarianism, and the coming (or presence) of war. Nevertheless, the decade is remembered in different ways in different parts of the world. For people in the United States, the 1930s was indelibly the age of the Great Depression. Bank panics destroyed faith in the economic system, and joblessness limited faith in the future. The worst drought in modern American history struck the Great Plains in 1934. Windstorms that stripped the topsoil from millions of acres turned the whole area into a vast Dust Bowl and destroyed crops and livestock in unprecedented amounts. As a result, some 2.5 million people fled the Plains states, many bound for California, where the promise of sunshine and a better life often collided with the reality of scarce, poorly paid work as migrant farm labourers.

For Americans, the 1930s will always summon up images of breadlines, apple sellers on street corners, shuttered factories, rural poverty, and so-called Hoovervilles (named for President Herbert Hoover), where homeless families sought refuge in shelters cobbled together from salvaged wood, cardboard, and tin. It was a time when thousands of teens became drifters; many marriages were postponed and engagements were interminable; birth rates declined; and children grew up quickly, often taking on adult responsibilities if not the role of comforter to their despondent parents. It was a time when the number of women in the workplace actually increased, which helped needy families but only added to the psychological strain on the American male, the traditional "breadwinner" of the American family. It was a time when one of the most popular tunes was "Brother, Can You Spare a Dime?"

GLOBAL CONCERNS

The memories of Europeans, by contrast, are haunted not by their economic difficulties, which were considerable, but by the spectre of Adolf Hitler and his drive to conquer the European continent. The Great Depression, of course, had created the perfect environment—political instability and an economically devastated and vulnerable populace—for the Nazi seizure of power and fascist empire building. Consequently, it was the spread of totalitarianism and not economic hardship that occupied the minds of Europeans in the 1930s. The situation was similar in Asia, where urban and rural penury was a normal feature of economic life; moreover, the decade of the 1930s is forever linked to the spread and brutality of Japanese imperialism. Thus, while Americans were preoccupied through most of the decade with their own domestic hardships, Europeans and Asians had other, more transnational, problems to confront.

The spread of totalitarianism

Moreover, the distinctive economic dilemmas of the 1930s were novel to Americans, largely because their historical experiences were so dissimilar from those of people in the rest of the world. For example, when British author George Orwell published *The Road to Wigan Pier* in 1937, he was describing an old problem: the class structure and its immemorial effect on workers in Britain. But when American authors such as Edmund Wilson or John Steinbeck wrote about the shut-down assembly lines in Detroit or the exodus of the Okies (Oklahomans displaced by the Dust Bowl) to California, they were describing something new: the near-total breakdown of a previously affluent economy. Americans were absorbed by their "Great Depression" because they had never before encountered such a widespread economic failure. This is why they, unlike their foreign counterparts, did not even begin to think about the approach of war or the dangers of totalitarianism until the end of the 1930s.

But no matter how insular Americans were through much of the decade, the world arrived on their shores in the 1930s. At the moment that Americans were worrying about their economy, European intellectuals, scientists, scholars, artists, and filmmakers were literally running for their lives. Where a lot of them ran was to the United States.

The most important event in the history of European culture in the 1930s was this massive hemorrhage of talent. No one was more responsible for transforming the cultural balance of power between Europe and the United States than Hitler. From the moment he assumed power in Germany in 1933, his book burnings, his firing of Jewish scholars in German universities, his assault on modern art, and his conquest of Europe at the end of the decade forced the most illustrious members of the European intelligentsia to flee, many of them first to France, then to the United States. Even a partial roster of émigrés to America in the 1930s is extraordinary. Among the natural scientists (most of whom were instrumental in constructing the atom bomb) were Albert Einstein, Enrico Fermi, Edward Teller, Leo Szilard, and Hans Bethe. The social scientists included Erik Erikson, Hannah Arendt, Erich Fromm, Paul Lazarsfeld, and Theodor Adorno. Philosophers such as Paul Tillich and Herbert Marcuse also emigrated, as did novelists and playwrights such as Thomas Mann, Vladimir Nabokov, and Bertolt Brecht. Musicians and composers included Igor Stravinsky, Béla Bartók, Arnold Schoenberg, Paul Hindemith, and Kurt Weill. Among the architects were Walter Gropius and Ludwig Mies van der Rohe. Painters and sculptors left too, notably Marc Chagall, Piet Mondrian, and Marcel Duchamp. And among those who found a home in (and helped to change) Hollywood were Fritz Lang and Billy Wilder—not to mention the Hungarian director Michael Curtiz, whose legendary *Casablanca* (1942) was in part a tribute to European refugee actors from Peter Lorre to Ingrid Bergman.

Emigration of European intelligentsia

Notably, not all persons seeking entry to the United States as refugees from Hitler's Germany were outstanding scholars, artists, scientists, or musicians. Most were average Europeans, but throughout the 1930s Congress chose not to liberalize the immigration laws to allow for more than the minimum quota of arrivals.

As a result of the massive intellectual and artistic emigration, by the end of the 1930s New York City and Hollywood had replaced Paris and Vienna as the home of Western culture—just as Washington would replace London and Berlin as the centre of Western politics and diplomacy at the end of World War II. To comprehend the America that became a postwar superpower, culturally as well as politically, it is necessary to understand how the United States responded to and emerged from its own singular experiences of the Great Depression in the 1930s.

POLITICAL MOVEMENTS AND SOCIAL CHANGE

Aside from the Civil War, the Great Depression was the gravest crisis in American history. Just as in the Civil War, the United States appeared—at least at the start of the 1930s—to be falling apart. But for all the turbulence and the panic, the ultimate effects of the Great Depression were less revolutionary than reassuring.

This was undeniably an era of extraordinary political innovation, much of it expressed in the reforms enacted by Franklin D. Roosevelt's New Deal and his administration's attempts to cope with the problems of poverty, unemployment, and the disintegration of the American economy. It was also a time when a significant number of Americans flirted with Marxist movements and ideas, as well as with the notion that the model for a more humane society could be found in the Soviet Union. Above all, it was a decade of cultural ferment, in which American writers, artists, and intellectuals experimented with new, more socially oriented forms of literature, painting, theatre, music, and mass entertainment.

Yet, paradoxically, the turmoil of the 1930s turned out to be predominantly conservative in its impact on American society. The Great Depression taught people of all social classes the value of economic security and the need to endure and survive hard times rather than to take risks with one's life or money. Moreover, faced with the spectre of totalitarian ideologies in Europe and Japan, Americans rediscovered the virtues of democracy and the essential decency of the ordinary citizen—the near-mythical "common man" who was celebrated in Roosevelt's speeches, Frank Capra's movies, and Norman Rockwell's paintings. Thus, a decade marked by fundamental—even radical—social change ended for most with a reaffirmation of America's cultural past and its traditional political ideals.

By contrast, many American intellectuals in the 1920s, disillusioned by what they considered the pointless carnage of World War I, had shown little interest in politics or social movements. Nor did they display much affection for life in the United States. Indeed, most American novelists, poets, artists, composers, and scientists continued to believe, as they had since the 19th century, that the United States was culturally inferior to Europe. So, to learn the latest modernist techniques in literature, painting, or music or to study the most advanced theories in physics or psychoanalysis, they assumed they had to go to London, Paris, Berlin, Vienna, or Copenhagen.

But the stock market crash in 1929, the factory closures and spiraling unemployment of the early 1930s, and Hitler's takeover of the German government in 1933 forced many "expatriates" not only to return to the United States but to become politically engaged in their home country. During the worst years of the Great Depression, between 1930 and 1935, this engagement often took the form of an attraction to Marxism, the Soviet Union, and the American Communist Party.

The appeal of Marxism

Marxism seemed to explain persuasively the causes of capitalism's collapse while also providing a vision of an alternative social order. The Soviet Union, the site of the first successful Marxist-inspired revolution, appeared by the 1930s to be a concrete embodiment of what many writers called (in characteristically pragmatic American terms) the socialist "experiment." In addition, from 1934 to 1939, the Soviet Union was the most uncompromising opponent of Nazi Germany, seeking alliances with Britain, France, and the United States and promoting a "popular front" partnership of liberals and socialists within the Western democracies to halt the spread of fascism in Europe and throughout the world. Nowhere did Moscow's desire for a broad antifascist coalition appear more genuine than in the Spanish Civil War (1936–39), when the Soviet Union was the only country besides Mexico to aid in any serious way the Spanish Republicans against the armies of Francisco Franco (supported by Hitler and Benito Mussolini).

Meanwhile, the communist parties in the United States and in western Europe gave intellectuals—as well as teachers, lawyers, architects, and other middle-class professionals—a feeling that they were no longer solitary individuals suffering from the failures of capitalism but belonged instead to a vibrant community of like-minded souls, in that they were participants in an international movement larger than themselves and that they were literally making history. For all these reasons Marxism, the Soviet Union, and the various national communist parties enjoyed a prestige and a popularity through much of the 1930s that they had never possessed in the 1920s and would never again enjoy after the Great Depression.

In 1932 the appeal of Marxism led 53 prominent American writers—including the novelists Sherwood Anderson and John Dos Passos, poet Langston Hughes, literary critics Edmund Wilson and Malcolm Cowley, philosopher Sidney Hook, and journalist Lincoln Steffens—to announce their support for William Z. Foster, the Communist Party's candidate for president. Although Dos Passos, Wilson, and Hook later became bitter critics of the Soviet Union's Stalinist regime (see STALINISM in the *Micropædia*), their initial enthusiasm for a socialist revolution indicated how compelling for intellectuals were the values and ideas of the left.

Hemingway's social commitment

Perhaps no writer better reflected this new sense of social commitment than Ernest Hemingway. In 1929 Hemingway published *A Farewell to Arms*. The novel's Lieutenant Henry, like Hemingway himself a volunteer American ambulance driver in Italy during World War I, decides to flee the madness of the war and make a "separate peace." Here, desertion is seen as an act of sanity, even of heroism. Eleven years later, in 1940, Hemingway published another novel about war—in this case, the Spanish Civil War—called *For Whom the Bell Tolls* (the title was taken from John Donne's poem, which is itself a hymn to human fellowship). In this novel, Robert Jordan, another Hemingwayesque volunteer, serving with a band of anti-Franco guerrillas, is badly wounded but stays behind to defend a bridge, thereby protecting his comrades as they retreat. Jordan—unlike Lieutenant Henry—has found a cause worth fighting and dying for. And Hemingway's own strong identification with the Spanish Republicans, for whom he raised money and helped make a documentary film called *The Spanish Earth* (1937), was symptomatic of a political involvement that neither he nor his fictional characters would have undertaken a decade earlier.

Of course, not every Depression-era American writer was entranced by communism or the Soviet Union. The majority of intellectuals and artists, like their fellow citizens, were much more comfortable voting for Roosevelt than idolizing Joseph Stalin. Indeed, by the middle and late 1930s, a growing number of American intellectuals—many of them clustered around the literary and political journal *Partisan Review*—had become militantly anti-Stalinist even as they retained their sympathy for socialism, their new stance having formed as Stalin launched a series of show "trials" that sent his former Bolshevik colleagues to Siberian labour camps (or more frequently to their death in the cellars of prisons), as terror spread throughout the Soviet Union, and as stories began to circulate about the communists murdering Trotskyists and anarchists behind the Republican lines in Spain. Still, it was not until August 1939, when Stalin shocked the world by signing a nonaggression pact with his archenemy Hitler that the Soviet Union and the Communist Party in the United States lost what was left of their moral authority with all but a few American intellectuals.

NEW FORMS OF CULTURAL EXPRESSION

The documentary impulse. Novelists, poets, painters, and playwrights of the 1930s did not need to be Marxists to create works that dealt with the problems of the Great Depression or the dangers of fascism. Indeed, even many who were sympathetic to Marxism acted as "fellow travelers" without joining the Communist Party. Most writers and artists in the prosperous 1920s thought of themselves as members of a transatlantic avant-garde and as stylistic disciples of Pablo Picasso, James Joyce, or Igor Stravinsky. In the impoverished and desperate 1930s, they repudiated—as did Malcolm Cowley in his literary memoir of the 1920s, *Exile's Return* (1934)—what they now regarded as the escapism and self-indulgence of their modernist mentors. Given the political and economic calamities at home and abroad, they sought to focus on the plight of workers, sharecroppers, African Americans, the poor, and the dispossessed. Further, they wanted to communicate their insights in a language—whether literary, visual, or musical—that their audiences could easily comprehend.

This impulse led, in a variety of genres, to an aesthetic of documentary-style realism and of social protest. For writers such as Edmund Wilson, Sherwood Anderson, John

Dos Passos, Erskine Caldwell, Richard Wright, and James Agee, fiction seemed inadequate in describing the disastrous effects of the Great Depression on political institutions, the natural environment, and human lives. So they joined with photographers and turned to journalism, as if their eyewitness portraits of desolate factories and American shantytowns, interviews with migrant workers and tenant farmers, and ubiquitous cameras could capture the "feel" and the essential truth of the Great Depression. Their yearning to record the pure, unadorned facts of daily existence, to listen to what Americans said about their plight, and to refrain from abstract theories or artistic embellishment was reflected in the titles of some of the books they wrote about their travels throughout the country: Wilson's *The American Jitters* (1932), Anderson's *Puzzled America* (1935), Nathan Asch's *The Road—In Search of America* (1937), Caldwell's *You Have Seen Their Faces* (1937), and Wright's *Twelve Million Black Voices* (1941).

Let Us Now Praise Famous Men

The most lyrical, and certainly the most eccentric, of these documentaries was *Let Us Now Praise Famous Men* (1941), with a text by Agee and pictures by Walker Evans. In order to illuminate the suffering but also the dignity of three sharecropper families in Alabama, Evans tried to photograph his subjects as objectively and as unobtrusively as possible. Meanwhile, Agee employed a variety of journalistic and artistic techniques: naturalistic description and dialogue, an almost anthropological itemization of clothing and household furniture, erudite discussions of agricultural problems in the Deep South, autobiographical ruminations, religious symbolism, and intimate expressions of love for the families and rage at their misery. Though the book's prose was perhaps too convoluted for readers in 1941, *Let Us Now Praise Famous Men* was the precursor of what would later be called the "new journalism," a highly personal style of reporting that influenced writers as diverse as George Orwell, Truman Capote, Tom Wolfe, and Norman Mailer.

Increasingly, Americans expected to be transported—through photos, newsreels, or radio—to the scene of the latest calamity. The urge to convey the sights and sounds of the 1930s was also reflected in the emergence of public-opinion polling as a major (if still primitive) industry, in the "living newspaper" productions of the Works Progress Administration (WPA) Federal Theatre Project that dramatized the headlines of the day, in government-sponsored documentary films such as Pare Lorentz's *The River* (1938) and *The Plow That Broke the Plains* (1936), in newsreels such as Twentieth Century Fox's Movietone News and Henry Luce's *March of Time*, in the photography of Dorothea Lange and Margaret Bourke-White, and in *Life* magazine's reliance on photographs even more than on traditional print journalism to tell the authentic story of what Americans were enduring at the time.

This sensation of being present, at least vicariously, at a crisis may explain why Orson Welles's radio adaptation on October 30, 1938, of H.G. Wells's *War of the Worlds* terrified so many listeners into believing that Martians had actually landed in New Jersey. The broadcast was done not as a play but in the style of a news story, with "announcers" breaking in for special bulletins, "reporters" delivering on-the-spot descriptions of the invasion, and "government spokesmen" (including one who sounded like FDR) issuing orders to troops and police. It was an event shared by millions of Americans, which is why it remains one of the most remembered events of the 1930s.

By the end of the decade, as Europe erupted into war, dramatic radio broadcasts took their cue from Welles's drama, and audiences grew to depend on a new type of foreign correspondent such as Edward R. Murrow, who broadcast from Berlin, Paris, or the rooftops of London and brought the sounds of falling bombs and air-raid sirens directly into people's living rooms, documenting a global struggle more cataclysmic than even Welles could have imagined.

Federal arts programs. The Roosevelt administration, too, embraced the notion that writers and artists should immerse themselves in the details, past and present, of American life. The United States, however, lacked a strong tradition of direct federal support for the arts. This may have been due to the public suspicion of such funding, especially during the 1930s, amid the spectacle of the Nazis' torchlight parades and their total control over radio and movies in Germany, which worried some U.S. congressmen and senators, as well as ordinary citizens, about the capacity of governments to use culture to manipulate public opinion. It was therefore both unprecedented and remarkable that between 1935 and 1939 the Roosevelt administration was able to create and sustain the Federal Art Project, the Federal Music Project, the Federal Writers' Project, and the Federal Theatre Project as part of the WPA.

The New Deal rationale for these cultural endeavours was that—just like construction workers—writers, musicians, painters, and actors had to eat and, more important, to use their skills for the benefit of society. Consequently, the Federal Theatre Project performances were staged not on Broadway but in working-class and African American neighbourhoods, outside factory gates, and in small towns whose residents had never seen a play. The Federal Writers' Project arranged for thousands of interviews with blue-collar workers, small farmers, fishermen, miners, lumberjacks, waitresses, and former slaves, and it published guidebooks that explored the history, ethnic composition, folklore, and ecology of every state. The Federal Music Project sponsored free concerts and the musical transcription of half-forgotten sea chanteys, cowboy and folk songs, Indian dances, Quaker hymns, and Negro spirituals. The Federal Art Project funded art education, established art centres, and made it possible for thousands of artists to complete works in sculpture, painting, and graphic arts; in addition, the Public Works of Art Project, influenced by Mexican painters such as José Clemente Orozco and Diego Rivera, arranged for murals to be painted on the walls of post offices and county courthouses depicting the stories of particular regions and local communities. It was precisely this attraction to traditional American melodies and to Norman Rockwell-like illustrations of ordinary life that helped composers such as Aaron Copland and Virgil Thomson and painters such as Thomas Hart Benton and Ben Shahn, all of them trained in the European modernist aesthetics of Stravinsky or Picasso, to adapt avant-garde techniques to "American" themes and hence offer an art accessible to popular taste.

The Federal Writers' Project

Theatre. None of this means that in the 1930s novelists abandoned fiction or that playwrights ignored the theatre. Rather, many writers still wanted to invest contemporary issues with poetic as well as political power, to raise brute facts to the level of art. Some, influenced by the Soviet Union's call for Socialist Realism, tried to create a didactic "proletarian" literature that usually chronicled a young, politically innocent worker's discovery of the need to join the labour movement, if not the Communist Party. This formula, with its melodramatic tale of how the exploited could triumph over the bosses, frequently led to wooden or bombastic prose, both in novels and on the stage.

Still, there were a number of theatrical companies in addition to the Federal Theatre—such as the Theatre Union and Orson Welles's Mercury Theatre—that attempted to put on plays that were artistically challenging as well as socially relevant. No company was more successful in this effort than the aptly named Group Theatre. Founded in 1931 by the directors Harold Clurman, Lee Strasberg, and Cheryl Crawford and featuring actors such as Stella Adler, John Garfield, Franchot Tone, Lee J. Cobb, Karl Malden, and Elia Kazan, the Group Theatre survived throughout the Great Depression in New York City as a noncommercial repertory company without stars or prima donnas, devoted to plays of current significance, and emphasizing a psychologically realistic acting style known as the Method, which Clurman and Strasberg borrowed from the ideas Konstantin Stanislavsky pioneered during his reign as director of the pre-Bolshevik Moscow Art Theatre.

The Group Theatre

In 1935 the Group's leading playwright, Clifford Odets, wrote a one-act play whose title could not have summed up more accurately the political sentiments of the 1930s: *Waiting for Lefty*. This was the quintessential proletarian drama in which the actors and the audience on opening night arose at the end of the play to demonstrate their solidarity with New York City taxi drivers by chanting "Strike! Strike! Strike!"

While some continue to see the Group's political engagement as its enduring hallmark, its true legacy lay not in its ideology but in its impact on American acting, especially on the screen. After World War II, under the influence of Strasberg, Adler, and Kazan, actors who trained in the Method—Marlon Brando, James Dean, Meryl Streep, Paul Newman, Robert De Niro, Al Pacino, Dustin Hoffman, and Shelley Winters, among others—became the most emotionally compelling performers in American movies.

Fiction. The social consciousness of the theatre was duplicated in some of the widely read novels of the 1930s. Here, too, authors strove for a fidelity to the sombre facts of the Depression experience. James T. Farrell's *Studs Lonigan* trilogy (1932, 1934, 1935) explored the claustrophobic world of lower-middle-class Irish Catholics, while Richard Wright's *Native Son* (1940) offered a harrowing portrait of a young African American man imprisoned in white America, capable of asserting his identity only through fear-drenched acts of violence.

It was this sense of constriction, the fear of shrinking natural and economic resources, the feeling that America was no longer buoyant and youthful—no longer a land of infinite hope and opportunity—that captured the mood of the 1930s and underlay the message of many of its novels. John Dos Passos's trilogy *U.S.A.* (1930, 1932, 1936)—a "multimedia history" of the United States in the first three decades of the 20th century, weaving together newspaper headlines, popular songs, biographies of celebrities, fictional stories, and eloquent prose-poems—was unrelenting in its sardonic depiction of American lives wasted in the neurotic pursuit of wealth and success. John Steinbeck's

The Grapes of Wrath

The Grapes of Wrath (1939), the most illustrious "protest" novel of the 1930s, was an epic tribute to the Okies, those throwbacks to America's 19th-century pioneers, now run off their farms by the banks, the Dust Bowl, and the mechanization of modern agriculture, clattering in their trucks and jalopies across the Arizona desert on Route 66 to the advertised promised land in California, a despised caste of migrant labourers who (like Steinbeck's heroic earth mother, Ma Joad) still insisted that the "people" are indestructible no matter what tragedies they must surmount.

But California might not have been a place for new beginnings; in the 1930s, as the novelist Nathanael West observed in *The Day of the Locust* (1939), it was more likely a destination where people went to die. In this novel, as well as in *Miss Lonelyhearts* (1933), West—in his fascination with bizarre personalities and psychological breakdowns—may well have expressed the deeper literary preoccupations of the 1930s more perceptively than did Wright or Steinbeck, preoccupations also reflected in John O'Hara's *Appointment in Samarra* (1934) and Horace McCoy's *They Shoot Horses, Don't They?* (1935).

Apolitical writers

Like West, the finest and most idiosyncratic writers of the decade—Thomas Wolfe, who was obsessed with dramatizing his own life in *Look Homeward, Angel* (1929); F. Scott Fitzgerald, whose *Tender Is the Night* (1934) and *The Last Tycoon* (1941) contained passages of prose as haunting as anything one could find in *The Great Gatsby* (1925); and William Faulkner, whose *The Sound and the Fury* (1929), *Light in August* (1932), and *Absalom, Absalom!* (1936) would appear on any list of the great American novels of the 20th century—did not conform to the formulas of protest or the demands of any creed. Their novels were not optimistic or pessimistic about America, nor were they "radical" or "conservative." More often, they were apolitical. Each of these authors strove not for a timely discussion of the social problems of the Great Depression years but for a timeless meditation on the agonies of life, love, and death. This sensitivity to private human predicaments, or more specifically to what might happen over a lifetime to husbands and wives and children in a small fictional New England village called Grover's Corners, was also why Thornton Wilder's *Our Town* (1939), not *Waiting for Lefty*, came to be the most treasured and enduring play of the 1930s. Such novels and plays—romantic, confessional, disturbing—would still be read or performed long after the proletarian aesthetic had lost its appeal for most Americans.

Popular culture. The indifference to politics and to the larger social concerns of the 1930s was reflected as well in the popular culture of the decade. In contrast to the prosperity of the Roaring Twenties, the 1930s emphasized simplicity and thrift. Although styles tended to reflect the glamour of contemporary movies, clothes themselves were mended before being replaced, and the invention of synthetic fibres led to the use of washable, practical, easy-care fabrics. Many who could not afford books or periodicals spent time reading in libraries. Inexpensive amusements included backyard games, puzzles, card games, and board games such as Monopoly, which was introduced in 1935. Even the national pastime, baseball, changed profoundly during the Great Depression. Major league rosters and players' salaries were cut, 14 minor leagues were eliminated, and, in an effort to bolster attendance that had fallen by more than 40 percent by 1933, night games were introduced. And with the end of Prohibition in 1933, nightclubs became legitimate places not only to consume liquor but to socialize, dance, enjoy the entertainment, and be seen wearing the latest fashions. Because radio and film reached many more people than novels or plays, some intellectuals believed that the mass media might be the most effective weapon for radicalizing Americans. Yet, predictably, the radio networks and the Hollywood studios, as commercial enterprises, were more interested in entertaining than in indoctrinating the masses.

Thus, the most popular programs on radio were afternoon soap operas, music and variety broadcasts, and half-hour comedy shows like *Amos 'n' Andy, The Jack Benny Program*, and *The Edgar Bergen/Charlie McCarthy Show*. Although Hollywood was filled with people sympathetic to the political left—people who frequently contributed money to the labour movement or the Spanish Republicans or who were indispensable in organizing the Screen Actors, Writers, and Directors guilds—little of this political activism left an imprint on the screen.

Hollywood's response

In fact, it is striking how few American movies during the 1930s dealt with the plight of the poor and the unemployed. The most memorable films of the decade (particularly those made at Metro-Goldwyn-Mayer, Paramount, and Twentieth Century Fox) were musicals, screwball comedies, and romances. Only Warner Brothers specialized in movies, usually gangster sagas, about the violence and poverty of slum life, a life the embattled hoodlum protagonists always yearned to escape.

What many of Hollywood's movies really had in common—even the spectacles of director Busby Berkeley and the dazzling duets of Fred Astaire and his frequent partner Ginger Rogers—was a soundtrack peppered with hard-boiled, even cynical, staccato chatter reminiscent of Walter Winchell's gossip columns in the newspapers and on the radio. The fast-talking guys and dames of 1930s movies—like the contemporaneous music and lyrics of George Gershwin and Ira Gershwin, Cole Porter, Irving Berlin, and Richard Rodgers and Lorenz Hart—were the product of a culture both urban and urbane; the movies and the music depended on clever allusions and witty dialogue, written or composed mostly by sophisticated Manhattanites. One could never imagine Cary Grant, Fred Astaire, Katharine Hepburn, Bette Davis, Rosalind Russell, Claudette Colbert, or the Marx Brothers portraying rural hayseeds or working stiffs. Nor was it possible to envision the gangsters, as played by Edward G. Robinson or James Cagney, asking passing strangers if they could spare a dime. The characters they played all lived in a world of posh furniture and polished floors, of well-cut suits and gowns, of elegant nightclubs filled with cigarette smoke and champagne and piano music, a world far removed from the one movie audiences inhabited.

Some of the music of the 1930s tried to assuage the social suffering. Indeed, from Lew Brown and Ray Henderson's "Life Is Just a Bowl of Cherries" (1931) to Al Dubin and Harry Warren's "We're in the Money" (1933), many of the era's popular songs were suffused in buoyant optimism. The emphatic "Happy Days Are Here Again" (1929) could be heard just about anywhere, whether as a political jingle for Roosevelt's 1932 presidential campaign or as the theme song for the *Your Hit Parade* radio show, launched in

1935. By mid-decade the Benny Goodman Orchestra had ushered in the swing era, popularizing a style of big band jazz that had been pioneered a decade earlier by African American ensembles led by Fletcher Henderson and Duke Ellington. Dance-oriented and relentlessly upbeat, swing was not a palliative for hopelessness; it was a tonic for recovery.

Yet songs that expressed a loss of faith in the American Dream were not completely absent. While Bing Crosby could sing "Just remember that the sunshine always follows the rain" in 1931 in "Wrap Your Troubles in Dreams," that same year he also recorded "Brother, Can You Spare a Dime?" Folk songs from the period, many recorded as part of the Federal Music Project's archival work, provide an especially vivid index of the deprivation suffered by ordinary Americans. Among the folksingers "discovered" through the field recordings of folklorists such as John Lomax and Alan Lomax was Leadbelly (Huddie Ledbetter), an ex-convict who gained fame for the songs he wrote about African American life during the Great Depression. No folk singer-songwriter, however, is more inextricably linked to the music of hardship and protest than Woody Guthrie. An Oklahoman, he took to the road at the height of the Dust Bowl era, frequenting hobo and migrant camps on his way to California, where he first popularized his songs about the plight of Dust Bowl refugees. With politically charged songs such as "(If You Ain't Got the) Do Re Mi," "Union Made," "Tom Joad" (inspired by *The Grapes of Wrath*), and "This Land Is Your Land," Guthrie became a mythic figure who continued his support of labour and radical politics (including his involvement with the Communist Party) long after most American intellectuals had abandoned them. In the process he became not only a catalyst for the folk music movement centred on New York City's Greenwich Village in the 1940s and '50s, with its strong association with leftist politics, but ultimately a role model for singer-songwriter Bob Dylan, who championed social protest in the early 1960s at the head of the folk music revival.

Woody Guthrie

In Hollywood, too, some of the leading directors of the 1930s, such as Capra in *Mr. Deeds Goes to Town* (1936) and *Mr. Smith Goes to Washington* (1939) or John Ford in his movie version of *The Grapes of Wrath* (1940), addressed the corruption of corporate and political power in modern America or the wretched conditions in which migrant farmers lived. The hollowed-out face of Henry Fonda as Steinbeck's Tom Joad, after all, was as potent an icon of the 1930s as Astaire's top hat and tails.

But few images from this period have lasted as long, or had as great an influence on filmmaking in America and abroad, as that of the fictional media mogul Charles Foster Kane in *Citizen Kane*. Directed by and starring a 25-year-old Welles and released in 1941, the movie was astonishing in part because of its stylistic virtuosity but also because it rebelled against the political clichés of the 1930s. By telling Kane's story from multiple perspectives, by presenting him as a man to be feared or pitied as well as occasionally admired, and by acknowledging at the end that no single word (not even "Rosebud") could explain a person's life, the movie refused to pass judgment or deliver a message—refused to say that this man of wealth and power is evil or that the society that produced him is in need of fundamental change. Neither sentimental nor propagandistic, *Citizen Kane* transcended the filmmaking conventions and the preconceptions of the 1930s and hinted at a more ironic age, with fewer certitudes, that would follow World War II.

PORTRAYALS OF HOPE

Americans in 1941, however, were not yet ready for the cool detachment of *Citizen Kane*. After 10 years of hard times, when the Depression felt like a natural as well as economic disaster (made worse by real environmental catastrophes such as floods and dust storms), what people wanted from their government and their popular culture was comfort. By the late 1930s, all but a few Americans longed not for revolution but for recovery, not for uncertainty but for stability, not for more social conflict but for a sense of national unity.

These essentially conservative impulses dominated the closing years of the Great Depression, though they had been present all along. Franklin D. Roosevelt recognized the craving for solace in the midst of chaos by clothing his reforms in conservative language. The very names of the New Deal agencies and programs—the National Recovery Administration, the Agricultural Adjustment Administration, the Civilian Conservation Corps, the Tennessee Valley Authority, Social Security—promised that America would be repaired and strengthened rather than transformed. Floods would be "controlled"; hydroelectric power would be "harnessed"; the soil would be "conserved"; order would be "restored." In short, Americans would get a new, fairer deal of the cards but not a brand-new game with perplexing new rules. Even African Americans—for many of whom the toils of the Great Depression were hardly different from the travails of everyday life in segregated America—found hope and inspiration in the New Deal, especially as it was enunciated by first lady Eleanor Roosevelt. They showed their support by switching their political allegiance from the Republican Party to the Democratic.

Conservative impulses

The Roosevelt administration's rhetoric and its policies were devised for a country that had shed the optimism and the innocence of the 1920s, a country that now regarded itself, psychologically, as middle-aged. The popular culture of the 1930s reinforced this perception that Americans had entered an era of limits, where they should make the best of what they already had rather than embarking on a quest for the unobtainable. The title of one of the decade's best-selling self-help books, *Life Begins at Forty* (1932) by Walter Pitkin, implied that a wise if chastened maturity was emotionally healthier and more realistic than adolescent self-confidence. At the same time, movies such as Capra's *It Happened One Night* (1934), *You Can't Take It with You* (1938), and *Meet John Doe* (1941) suggested that people were happier and better off if they were not rich and that the familiar pleasures of home and family were more fulfilling than the ambitions of the powerful or the affectations of the elite. This was a soothing idea for people whose dreams of a more affluent and adventurous life had vanished.

The conservatism of the 1930s coincided with a revival of interest in the American past and a veneration of America's legendary heroes. The publication of multivolume biographies of George Washington, Andrew Jackson, and Robert E. Lee, or epic poems like Archibald MacLeish's *The Land of the Free* (1938), reminded people of the leaders (whatever their differing philosophies) who had guided the nation through its earlier crises. This reverence for tradition, which encouraged Americans to believe they could prevail over their current predicaments, was the subtext of the decade's most famous novel and the movie that set box-office records both in the 1930s and for the next half century, *Gone with the Wind* (book, 1936; movie, 1939).

The resurgence of cultural nationalism was hardly unique to the United States. Britain, France, Germany, Italy, the Soviet Union, and Japan were all competing with one another in the glorification of their histories and their values through international automobile races, aviation speed and endurance contests, the acquisition of gold medals at the Olympic Games of 1932 and 1936, and shortwave overseas radio broadcasts such as the British Broadcasting Company's Empire Service. By the end of the 1930s, the Roosevelt administration—fearing the spreading influence of Germany and Italy through the growth, in Latin America, of large émigré populations from those two countries—had entered the culture war by establishing libraries, educational exchanges, and U.S. schools in Mexico, Brazil, Argentina, and Chile. These initiatives marked the beginning of the U.S. government's far more extensive strategy of exporting American culture as an instrument of foreign policy during World War II and the Cold War.

Resurgence of cultural nationalism

Even as the worst economic problems of the Great Depression began to lift, the prevailing mindset could not forget the lessons of the era. The trust in the federal government to solve or at least address the fundamental dilemmas of various groups in American society (the elderly with Social Security, blue-collar workers with the Na-

tional Labor Relations Act, poor Southern farmers with the Tennessee Valley Authority), the dependence on Washington as the ultimate supervisor of corporate behaviour, the thirst for social and psychological security, the need to hold a job and save money as protection against some future economic crisis—all of these predilections continued to shape the mentality of Americans who lived through the Great Depression even after America's victory in World War II and the return of prosperity. The emotional scars, the fear of fear itself, could never be eradicated.

But the Great Depression and its aftermath also encouraged a faith in, and a love of, what America presumably stood for. These were not the sort of feelings one might have expected in a decade in which many people were initially angry about the failure of America's economic and social arrangements. Yet the transition from rage to reconciliation was reflected, symbolically, in one of the decade's most cherished movies, *The Wizard of Oz* (1939). Here Dorothy (played by Judy Garland) is transported from her drab, gray Kansas farm to the magical and Technicolor land of Oz. She and her companions—a scarecrow, a tin woodsman, and a cowardly lion—each seeking to change themselves or their circumstances, go off to see the wizard "because of the wonderful things he does." Although the wizard turns out to be a charlatan, he has an important lesson to teach, not just to his supplicants but to audiences in the 1930s. People, he says, do not need a wizard and his miracles; all they need to do is look inside themselves. So a movie that begins with Dorothy imagining a fantasy world somewhere over the rainbow ends with her back in Kansas, proclaiming "There's no place like home."

And in the midst of World War II, as the economy recuperated and people went back to work, the virtues of life at home became more palpable to most Americans. In 1939 John Steinbeck had portrayed an Oklahoma in *The Grapes of Wrath* that, like the rest of America, was still marked by scarcity and deprivation. In 1943 Richard Rodgers and Oscar Hammerstein opened a new musical, called *Oklahoma!*, on Broadway. Their Oklahoma, unlike Steinbeck's Dust Bowl, was a bountiful land where the corn was as high "as an elephant's eye."

The musical, with its joyous evocation of beautiful mornings, summed up the spirit of a people who had finally freed themselves from the constraints of the 1930s and could once again relish the vitality of the United States. It was this America—having survived its idiosyncratic crisis in the 1930s and having escaped the bombing of its cities and the destruction of its natural resources during World War II—that the rest of the world would have to decipher and deal with in the postwar years. (R.H.Pe.)

BIBLIOGRAPHY

Economic history. MILTON FRIEDMAN and ANNA JACOBSON SCHWARTZ, *A Monetary History of the United States, 1867–1960* (1963, reissued 1993), chapter 7, "The Great Contraction," is the single most important study of the Great Depression in the United States, detailing ways in which banking panics and monetary contraction contributed to the economic downturn.

Scholarly studies that analyze the role of particular factors in the American Depression include BEN S. BERNANKE, "Nonmonetary Effects of the Financial Crisis in the Propagation of the Great Depression," *American Economic Review*, 73(3):257–276 (June 1983); STEPHEN G. CECCHETTI, "Prices During the Great Depression: Was the Deflation of 1930–1932 Really Unanticipated?," *American Economic Review* 82(1):141–156 (March 1992); CHRISTINA D. ROMER, "The Great Crash and the Onset of the Great Depression," *Quarterly Journal of Economics*, 105(3):597–624 (August 1990); and PETER TEMIN, *Did Monetary Forces Cause the Great Depression?* (1976). JOHN KENNETH GALBRAITH, *The Great Crash, 1929* (1954, reissued 1997), is a riveting account of the 1929 stock market crash, one of the events leading up to the Great Depression in the United States.

International economy: BARRY EICHENGREEN, *Golden Fetters: The Gold Standard and the Great Depression, 1919–1939* (1992, reissued 1995), is an important study of the functioning and effects of the international gold standard in the interwar era. W. ARTHUR LEWIS, *Economic Survey, 1919–1939* (1949, reissued 1969), while somewhat dated, is an exceedingly useful survey of the nature and causes of the Depression in Great Britain, Germany, Russia, France, Japan, and the United States. Other works analyzing the Depression outside the United States include HAROLD JAMES, *The German Slump: Politics and Economics, 1924–1936* (1986); CHARLES P. KINDLEBERGER, *The World in Depression, 1929–1939*, rev. and enlarged ed. (1986); and ROSEMARY THORP (ed.), *Latin America in the 1930s: The Role of the Periphery in World Crisis* (1984).

Recovery: LESTER V. CHANDLER, *America's Greatest Depression, 1929–1941* (1970), provides a detailed description of the many programs implemented to deal with the Depression in the United States. BARRY EICHENGREEN and JEFFREY SACHS, "Exchange Rates and Economic Recovery in the 1930s," *Journal of Economic History*, 45(4):925–946 (December 1985), discusses how devaluation and monetary expansion contributed to economic recovery from the Depression in many countries. Two studies that examine the role of policy in ending the American Depression are E. CARY BROWN, "Fiscal Policy in the 'Thirties: A Reappraisal," *American Economic Review*, 46(5):857–879 (December 1956); and CHRISTINA D. ROMER, "What Ended the Great Depression?," *Journal of Economic History*, 52(4):757–784 (December 1992).

Impact: MICHAEL D. BORDO, CLAUDIA GOLDIN, and EUGENE N. WHITE (eds.), *The Defining Moment: The Great Depression and the American Economy in the Twentieth Century* (1998), includes a series of papers by distinguished scholars on the long-run impact of the Great Depression in the United States. JOHN MAYNARD KEYNES, *The General Theory of Employment, Interest, and Money* (1936, reissued 1997), is the pathbreaking work of economic theory that was inspired by the Great Depression and led to the rise of stabilization policy in the postwar era.

(C.D.Ro.)

Culture and society in the Great Depression. T.H. WATKINS, *The Hungry Years: A Narrative History of the Great Depression in America* (1999), is a comprehensive political and social history of the Great Depression in the United States; while PIERS BRENDON, *The Dark Valley: A Panorama of the 1930s* (2000), takes a more international approach, comparing the effects of the Depression in the United States, Britain, Germany, France, the Soviet Union, China, and Japan.

Other works that deal with cultural issues, both in the 1930s and in the 20th century, include RICHARD H. PELLS, *Radical Visions and American Dreams: Culture and Social Thought in the Depression Years* (1973, reprinted 1998); WARREN I. SUSMAN, *Culture as History: The Transformation of American Society in the Twentieth Century* (1973, reissued 2003); LAWRENCE W. LEVINE, *The Unpredictable Past: Explorations in American Cultural History* (1993); and MICHAEL KAMMEN, *American Culture, American Tastes: Social Change and the 20th Century* (1999). Chapter 1 of RICHARD H. PELLS, *Not Like Us: How Europeans Have Loved, Hated, and Transformed American Culture Since World War II* (1997), explores the role of American foundations in bringing refugee scholars, scientists, artists, and filmmakers to the United States in the 1930s, and it discusses the Roosevelt administration's efforts to export U.S. culture to Latin America at the end of the decade.

Two memoirs are still useful in illuminating the cultural and intellectual preoccupations of the 1930s: HAROLD CLURMAN, *The Fervent Years: The Story of the Group Theatre and the Thirties* (1945, reprinted 1983), is a personal history of the Group Theatre by one of its founders; and ALFRED KAZIN, *Starting Out in the Thirties* (1965, reprinted 1989), describes the polemical battles on the left and explores American reactions to the announcement of the Nazi-Soviet Pact in 1939. MARIA DIBATTISTA, *Fast-Talking Dames* (2001, reissued 2003), is excellent on the movies of the 1930s and on the actresses who delivered the witty dialogue that was Hollywood's trademark during these years. (R.H.Pe.)

Diagnosis and Therapeutics

Diagnosis, from the Greek *gnosis* meaning knowledge, is the art of determining the nature of a disease and distinguishing one disease from another. The diagnostic process is the method by which health professionals identify one disease rather than another as the most likely cause of a person's symptoms. Symptoms that appear early in the course of a disease are often more vague and undifferentiated than those that arise as the disease progresses, making the early stages the most difficult time to make an accurate diagnosis. Reaching an accurate conclusion depends on the timing and the sequence of the symptoms, past medical history and risk factors for certain diseases, and a recent exposure to disease. The physician, in making a diagnosis, also relies on various other clues such as physical signs, nonverbal signals of distress, and the results of selected laboratory and radiological tests. From the large number of facts obtained, a list of possible diagnoses can be determined, which are referred to as the differential diagnosis. The physician organizes the list with the most likely diagnosis given first. Additional information is identified, and appropriate tests are selected that will narrow the list or confirm one of the possible diseases.

Therapeutics is the art and science of treating disease. It comes from the Greek *therapeutikos,* which means "inclined to serve." In a broad sense therapeutics means serving and caring for the patient in a comprehensive manner, preventing disease as well as managing specific problems. Exercise, diet, and mental factors are therefore integral to the prevention as well as the management of disease processes. More specific measures that are employed to treat specific symptoms include the use of drugs to relieve pain or treat infection, surgery to remove diseased tissue or replace poorly functioning or nonfunctioning organs with fully operating ones, and counseling or psychotherapy to relieve emotional distress caused by the illness or its treatment. Confidence in the physician and in the method selected enhances effectiveness.

For coverage of related topics in the *Macropædia* and *Micropædia,* see the *Propædia,* Part Four, Division II, especially sections 423 and 424.

This article is divided into the following sections:

Diagnosis 246
- Historical aspects 246
- Medical history 247
- Physical examination 248
  - Manual procedures
    - Inspection
    - Palpation
    - Percussion
    - Auscultation
  - Special examinations
- Tests and diagnostic procedures 250
  - Clinical laboratory tests
  - Genetic testing
  - Instrumental screening
  - Surgical examination
  - Radiological screening
  - Computerized body scanning
- Formulating a diagnosis 256

Therapeutics 256
- Preventive medicine 256
- Treatment of symptoms 257
  - Pain
  - Nausea and vomiting
  - Diarrhea
  - Cough
  - Insomnia
- Designing a therapeutic regimen 258
- Diet 259
  - Prophylactic measures of nutrition
  - Therapeutic measures of nutrition
- Biological therapy 260
  - Blood and blood cells
  - Plasma
  - Immunoglobulins
  - Bone marrow transplantation
  - Hematopoietic growth factors
  - Biological response modifiers
  - Hormones
- Drug therapy 262
  - General features
  - Systemic drug therapy
  - Local drug therapy
  - Chemotherapy
- Surgical therapy 265
  - Major categories of surgery
  - Surgical techniques
- Radiation and other nonsurgical therapies 267
  - Radiation therapy
  - Other noninvasive therapies

Bibliography 268

DIAGNOSIS

Historical aspects

Traditionally, diagnosis has been defined as the art of identifying a disease from its signs and symptoms. Formerly, few diagnostic tests were available to assist the physician, who depended on medical history, observation, and examination. Only recently, with the many technological advances in medicine, have tests become available to assist in making specific diagnoses.

Medicine and personal hygiene reached new heights in the 5th century BC at the time of the Greek physician, Hippocrates. The Greeks recognized the salubrious effects of bathing, fresh air, a good diet, and exercise, which have received renewed attention today. Illness was thought to result from an imbalance between the four humours of the body: blood, phlegm, yellow bile, and black bile. The Greeks emphasized the value of observation, including bodily signs and excretions. The focus, however, was more on predicting the outcome of an illness (*i.e.,* prognosis) and less on its diagnosis. A physician's reputation depended on accurate prognostic skills, predicting who would recover and who would die or how long an illness would last.

The Greeks

Hippocrates is credited with establishing the ethical basis of the physician's behaviour, and graduating physicians still recite the oath ascribed to him. His writings document the value of objectively evaluating all aspects of the patient's symptoms, diet, sleep patterns, and habits. No finding was considered insignificant, and physicians were encouraged to employ all their senses—sight, hearing, smell, taste, and touch—in making a diagnosis. These principles hold just as true today.

The Romans made significant advances in supplying and purifying water and in improving sanitation.

Galen (AD 130–200) is considered the most influential physician after Hippocrates because of his extensive studies in anatomy and physiology. His voluminous writings in anatomy and physiology rendered him the ultimate authority in these fields until the 16th century. As the first

experimental neurologist, he described the cranial nerves and the sympathetic nervous system. He showed that the heart will continue beating when removed from the body and thus does not depend on the nervous system. Many of his views, however, contained fallacies, which remained unchallenged for centuries. His description of the heart and its chambers and valves, in which he contended that blood passes from the right to the left ventricle by means of invisible pores in the interventricular septum, delayed the discovery of blood circulation for 14 centuries. The true nature of the circulation of blood was not recognized until the time of William Harvey (1578–1657), who published his findings in *Exercitatio anatomica de motu cordis et sanguinis in animalibus* (translated as *An Anatomical Dissertation Upon the Movement of the Heart and Blood in Animals* and usually referred to as *De Motu Cordis*).

From the Middle Ages to the 18th century, uroscopy (examination of the urine) was a common method for diagnosing illness. The colour of the urine, as well as cloudiness, precipitates, and particles in the urine, was believed to indicate the cause of the disorder.

Diagnostic tools

One of the greatest advances in diagnostic tools was the invention of the compound microscope toward the end of the 16th century by the Dutch spectacle makers Hans Jansen and his son Zacharias. In the early 17th century Galileo constructed a microscope and a telescope. One of the great early microscopists, Antonie van Leeuwenhoek (1632–1723), was the first to see protozoa and bacteria and the first to describe red blood cells. He also demonstrated the capillary anastomosis (network) between arteries and veins that proved Harvey to be correct.

Although the mercury thermometer of Daniel Fahrenheit (1686–1736) appeared about 1714, it was not until 1866 that it came into general use as a clinical tool. It was initially 25.4 centimetres (10 inches) long and took five minutes to register a temperature. A pocket version was developed by Sir Thomas Clifford Allbutt in 1866. The thermometer was popularized by Karl August Wunderlich who thought, incorrectly, that every disease had its own characteristic fever pattern.

Another significant medical advance, which greatly improved the ability to diagnose diseases of the chest and heart, was the invention of the stethoscope in 1816 by René-Théophile-Hyacinthe Laënnec (1781–1826). Before this, the lungs and heart were examined by applying the ear to the chest wall. Laënnec initially used a roll of papers to enhance sounds from the chest and later replaced this "instrument" with a wooden cylinder. He improved the original monaural (one-ear) stethoscope with the binaural device still in current use. Tuberculosis was prominent at the time, and the stethoscope allowed Laënnec to diagnose this condition at an earlier stage than was previously possible.

Another significant diagnostic aid was the ophthalmoscope developed by Hermann von Helmholtz (1821–94), a physician best known for his knowledge of physics and mathematics. With this device, the retina and blood vessels could be seen through the pupil, allowing the inner eye to provide information not only concerning diseases of the eye but also about those pertaining to cardiovascular abnormalities and complications of diabetes mellitus.

Perhaps the greatest modern anatomic diagnostic tool is the X ray, discovered in 1895 by the German physicist Wilhelm Conrad Röntgen. X rays have since been commonly referred to as roentgen rays, and their application eventually led to the development of computerized tomography and magnetic resonance imaging, two techniques that are extremely useful modern diagnostic tools.

Medical training

The training of physicians also has undergone significant change over the years. Until the end of the 19th century, physicians were trained through lectures and rarely were taught at the patient's bedside. This practice was altered by Sir William Osler, one of the most renowned physicians of the early 20th century, who introduced the practice of instructing students at the bedside of the patient. He emphasized the importance of taking an accurate medical history, providing a thorough examination of a patient, and closely observing the patient's behaviour to gather clues for a diagnosis before resorting to laboratory testing.

Medical history

The medical history of a patient is the most useful and important element in making an accurate diagnosis, much more valuable than either physical examinations or diagnostic tests. The medical interview is the process of gathering data that will lead to an understanding of the disease and the underlying physiological process. To be effective, an interviewer must possess good communication skills and be alert to nonverbal clues as well as to the verbal message. Often, more information is conveyed by nonverbal actions and tone of voice than by words. The objective is to obtain an accurate and comprehensive picture of the patient's situation, including the nature and timing of symptoms, emotional factors (including types of stress), and past medical conditions that may place the patient at greater risk for certain diseases.

The medical interview

The accuracy and usefulness of the medical interview depend on the physician's ability to elicit information pertinent to the problem at hand and on the patient's accurate recall and articulation of the sequence of symptoms. This may be difficult because meaningful data may be forgotten if the patient is experiencing pain or emotional distress. The skilled interviewer knows when to use silence, open-ended questions, or specific closed-ended questions to explore avenues in which the most useful information may be found. The real reason for the patient's visit may not be apparent until a rapport has been established and the person feels comfortable describing what is most bothersome. Problems that are emotionally threatening may not be voiced until adequate courage is summoned—sometimes not until the end of the appointment when the patient's hand is on the doorknob.

A complete medical history consists of an account of the present illness, the past medical history, family history, occupational background, psychosocial history, and a review of body systems.

An account of the present illness, which includes the circumstances surrounding the onset of recent health changes and the chronology of subsequent events that have led the patient to seek medical care, is essential to understanding the course of the disease process. Current medications are listed in the medical history because they may play a role in the current illness.

The past medical history is an overall view of the patient's health prior to the present illness. It should include previous hospitalizations, injuries, operations, and any significant illness that may not have required hospitalization. Allergies are included here if not listed separately.

Family history

Included in a family history are the age and state of health of each immediate family member as well as the cause of death of any parents, grandparents, and other close relatives. Of particular importance are genetic or environmental diseases that have known risks. If a close relative such as a father died of a heart attack (acute myocardial infarction) before the age of 60, all his children are at greater risk of suffering an early heart attack. This risk increases if other factors such as hypertension (high blood pressure) or elevated serum cholesterol are present. Similarly, a history of some cancers (*e.g.*, colon cancer) increases the risk for offspring to develop that type of cancer. The development of lung cancer in a parent provides even greater impetus for close relatives to avoid smoking. Other diseases that may have hereditary or environmental roots are diabetes, hypertension, tuberculosis, depression and other forms of mental illness, arthritis, and epilepsy. Actually, any disease that arises in two or more members of a family suggests a possible predisposing factor, and the patient should be considered to be at increased risk for this condition.

The occupational history is important because the workplace may be a source of toxins, such as chemicals or cigarette smoke, that place one at higher risk of cancer or other diseases.

The psychosocial history—information on education, lifestyle, marital status, and religious beliefs—may influence future medical decisions, as may the patient's smoking history, alcohol intake, and use of such controlled substances as marijuana or cocaine.

The review of body systems allows the physician to identify any other symptoms that have not been noted previously and that may influence the patient's current state of health or provide subtle clues to the diagnosis. All major body systems are reviewed in an orderly manner, usually from the head down to the extremities. The intent is to uncover any past illnesses or problems that have not been previously identified and that may now or later influence the patient's health. For example, the patient may describe leg pain while walking, which could be an early indication of blood vessel occlusion and increase the physician's concern about possible coronary artery disease that otherwise may not have been suspected.

Physical examination

MANUAL PROCEDURES

The physical examination continues the diagnostic process, adding information obtained by inspection, palpation, percussion, and auscultation (see below). When data accumulated from the history and physical examination are complete, a working diagnosis is established, and tests are selected that will help to retain or exclude that diagnosis.

Patients are usually apprehensive and anxious when being examined because they feel exposed, vulnerable, and afraid of discomfort. The physician attempts to allay that anxiety by explaining which examinations are to be performed and the degree of discomfort that will be entailed. Throughout the examination, concern for the patient's dignity must be maintained.

Inspection. A wide array of sophisticated instruments is available to assist with examinations, but a well-performed visual inspection can often reveal more information. Osler admonished physicians to closely observe patients before touching them, to cultivate the power of observation, as it is one of the greatest diagnostic tools. Wasting and hallmarks of poor nutrition may indicate chronic disease; poor grooming or slack posture may suggest depression or low self-esteem.

Visual inspection

Inspection begins with the patient's general appearance, state of nutrition, symmetry, and posture. The physician then proceeds to more specific examination of the skin—looking for redness or other signs of infection, hair loss, nail thickening, and moles or other areas of pigmentation—and inquires about any recent changes in skin lesions that could indicate early cancer.

Examination of the nails can provide important clues about systemic disease. Clubbing of the nails (broadening of the nailbeds, with curved and shiny nails) may indicate congenital heart disease, chronic obstructive pulmonary disease, bronchogenic carcinoma, or another cardiac or pulmonary condition. Pitting of the nails occurs in about 50 percent of patients with psoriasis.

Inspection should encompass, in particular, areas that the patient normally would not be able to see, such as the scalp, back, and buttocks.

The skin should always be inspected for cancer, though it is sometimes difficult to differentiate a benign mole (nevus) from a cancer. The most dangerous skin cancer, malignant melanoma, occurs in about 1 in 10,000 people and can spread readily throughout the body. A squamous-cell carcinoma also may spread but is slow to do so and can be completely cured by early detection and removal. Basal-cell cancer is the most common form of skin cancer, and, though it is locally invasive, it does not spread. Suspicious lesions are those that have recently enlarged, started to bleed, become darker, or developed an irregular outline. Most skin cancers occur on areas of the body that have been exposed to the sun; they are more common in light-skinned individuals with blond hair and blue eyes who sunburn easily.

The most common premalignant (precancerous) skin lesion is actinic keratosis, a rough, scaling, red or brown papule that appears on sun-exposed areas such as a bald scalp, ears, the forehead, and the back of the hands. These lesions can be easily removed by cryotherapy (therapeutic use of cold) or electrodesiccation (dehydration of tissue by electric current).

Palpation. Palpation is the act of feeling the surface of the body with the hands to determine the characteristics of the organs beneath the surface. It can be performed with one hand or two and can be light or deep.

Types of palpation

Light palpation is used to detect tenderness, muscle spasm, or rigidity of the abdomen. If abdominal pain is present, gentle palpation begins farthest away from the pain to localize the point of maximum tenderness. Acute inflammation in the abdomen, as in acute appendicitis, causes peritoneal irritation, resulting in not only localized tenderness in the right lower abdomen but also a guarding reaction (tightening and rigidity) by the muscles in that area to protect the inflamed organ from the external pressure.

Deep palpation of the abdomen is used to determine the size of the liver, spleen, or kidneys and to detect an abnormal mass. An abdominal aortic aneurysm can be detected by palpating a pulsatile mass in the upper abdomen. An acutely tender mass in the right upper abdomen that is more painful on inspiration is probably an inflamed gallbladder. An unexplained nontender abdominal mass could be as nonthreatening as a hard stool or as serious as a tumour.

Palpation also is used to detect and evaluate abnormal lesions in the breast, prostate, lymph nodes, or testicles. Proper breast examination includes frequent (at least monthly) self-examinations and an annual examination by a physician. Palpation should be methodical and performed over the entire breast; the method of action is done either in concentric circles or outward from the nipple using a spokes-of-a-wheel approach. Suspicious breast lesions are hard and fixed rather than movable. Skin retraction or breast asymmetry can indicate an underlying, potentially serious lesion. Cancers are usually not tender, and benign lesions are more likely to be round, elastic or firm, movable, and well-defined. Similarly, suspicious prostate lesions are hard, irregular nodules within the prostate, whereas benign prostatic hyperplasia (BPH) is a soft symmetrical enlargement of the gland.

Palpation also can detect cardiac enlargement if the point of maximal impulse (PMI) of the heart is farther left than normal. Other cardiac abnormalities can be suspected if a thrill is felt using light palpation over the chest wall. A thrill is a vibratory sensation felt on the skin overlying an area of turbulence and indicates a loud heart murmur usually caused by an incompetent heart valve.

Percussion. Percussion is a diagnostic procedure used to determine the density of a part by tapping the surface with short, sharp blows and evaluating the resulting sounds. In the abdomen it can be used to detect fluid (ascites), a gaseous distention of the intestine as occurs in bowel obstruction, or an enlargement of the liver. It is used most often to evaluate the chest. Percussion produces a resonant note when the area over a healthy lung is struck; a dull sound, however, will emanate if the lung contains fluid, as in pneumonia, or when a region over a solid mass such as the heart is tapped. A lung that is diseased with emphysema contains more air than a healthy lung and produces hyperresonance. A stomach distended with air will produce a high-pitched, hollow tympanic sound.

Auscultation. Auscultation is performed with a stethoscope to evaluate sounds produced by the heart, lungs, blood vessels, or bowels. The lack of bowel sounds indicates a nonfunctioning or paralyzed bowel, and high-pitched "tinkling" bowel sounds suggest bowel obstruction. The "growling" of the stomach is an accentuation of these sounds during periods of bowel hyperactivity.

Body sounds provide diagnostic clues

Bruits are blowing vascular sounds resembling heart murmurs that are perceived over partially occluded blood vessels. When detected over the carotid arteries, a bruit may indicate an increased risk of stroke; when produced by the abdomen, it may indicate partial obstruction of the aorta or other major arteries such as the renal, iliac, or femoral arteries.

Listening to the sound of air passing in and out of the lungs can be useful in detecting an obstruction as in asthma or an inflammation as in bronchitis or pneumonia. Adventitious sounds are those heard in addition to normal breathing sounds and include crackles, wheezes, and rubs. Crackles (also called rales) resemble the sound

made by rubbing hair between the fingers next to the ear. They are caused by fluid in the small passageways that adheres to the walls during respiration. Crackles are heard in congestive heart failure and pneumonia. Wheezes, musical sounds heard mostly during expiration, are caused by rapid airflow through a partially obstructed airway as in asthma or bronchitis. Pleural rubs sound like creaking leather and are caused by pleural surfaces roughened by inflammation moving against each other, which occurs in patients with pneumonia and pulmonary infarction.

Cardiac auscultation is the evaluation of the sounds made by the heart valves—namely, the aortic, pulmonary, tricuspid, and mitral—for murmurs that may be due to turbulent blood flow or vibrations from a heart valve deformity. Murmurs may be physiological (unimportant clinically) or pathological, indicating a problem that needs attention, especially if they reflect obstruction of normal blood flow. Heart murmurs vary according to their timing in the cardiac cycle (systole, the period of contraction when blood is pumped from the heart, and diastole, the period of heart expansion between pumping), location, duration, intensity, pitch, and quality. Intensity is graded on a scale from 1 to 6, with 6 being the loudest. Heart murmurs are described as "grade 2/6"—the numerator represents the intensity of the murmur, and the denominator indicates the highest grade of the scale being used. However, the intensity of the murmur alone provides little information about the clinical severity of the problem. An ejection murmur caused by turbulence across the aortic valve during systole can be either serious or nonthreatening depending on its cause, even though the intensity of the murmur may be the same. Therefore, the pitch and quality of the murmur also are described. Pitch is usually reported as high or low, and quality is described as harsh, soft, blowing, musical, or rumbling. For example, the murmur of mitral stenosis may be described as a grade 3/6, low-pitched, rumbling, presystolic murmur heard best at the apex and having an increased first heart sound at the apex.

SPECIAL EXAMINATIONS

Evaluation of body systems

Emergency. Of greatest importance in an emergency is the evaluation of systems that are essential to sustaining life—namely, the circulatory, respiratory, and central nervous systems.

First, the person in distress should be checked to determine whether breathing is normal or at least whether there is adequate exchange of air to ensure oxygenation of the blood. If the person is unconscious, the first maneuver is to tilt the head back and lift the chin (unless a neck injury is suspected) to prevent the tongue and jaw from obstructing the airway and then to provide artificial respiration. If the person has eaten recently, the cause may be obstruction by a foreign body (usually food), and the Heimlich maneuver (subdiaphragmatic abdominal thrust) should be performed.

Second, the circulation should be evaluated. Is the heart beating, and is the output adequate to provide oxygenated blood to the tissues, or has this been compromised by excessive bleeding? A blood pressure greater than 100/60 millimetres of mercury (mm Hg) indicates adequate perfusion.

Shock occurs when the blood pressure falls to extremely low levels because of inadequate blood volume (hypovolemic shock), poor heart function (cardiogenic shock), or malfunction of the vascular system that results in lost peripheral vascular tone, vasodilation, and pooling of the blood (neurogenic shock). Signs of shock are a rapid and weak pulse, pale complexion, sweating, and confusion. Organs particularly sensitive to injury if the shock is not corrected are the brain, heart, lungs, kidneys, and liver.

Examination of the unconscious patient

An unconscious person may not respond to external stimulation, in which case the person would be in a coma, or the patient may exhibit varying levels of unconsciousness, responding only to painful stimuli (deep level of unconsciousness) or when shaken or called by name (light level). Pupil size and reactivity to light can provide clues to the status of the nervous system. Bilateral dilated pupils that do not contract when a light is placed on one of them indicate death or severe damage to the nervous system. Small pupils that do react to light are seen in narcotic overdose. If one pupil is larger than the other, a brain lesion on one side or hemorrhage should be suspected.

Pediatric. Examinations to assess the well-being of children begin at birth. The Apgar score, named for the anesthesiologist Virginia Apgar, is obtained at one and five minutes after birth and indicates the condition of the newborn. A score of 0 (absent), 1, and 2 is given for each of the five parameters, which are heart rate, respiratory effort, muscle tone, reflex irritability, and colour. Infants scoring between 7 and 10 at one minute will likely do well with no special treatment; those scoring between 4 and 6 may require stimulation or brief respiratory support; those scoring 3 or below will probably need extended resuscitative efforts. Infants who have a score of 7 or above at five minutes will continue to do well. The Apgar score is usually reported as two numbers, from 1 to 10, that are separated by a virgule, the first number being the score at one minute, the second the score at five minutes.

Developmental assessment is measured with growth charts developed by the National Center for Health Statistics. A child's length (or height) and weight are plotted over time on standard graphs constructed from data gathered from a large number of average-sized children. The average length of a newborn infant is 50 centimetres (20 inches). The length increases by 50 percent at 12 months of age and doubles to 100 centimetres when the child is 4 years old. The average weight at birth is 3.4 kilograms (7.5 pounds), which doubles in 4 to 5 months and triples when the child is 12 months old. After 2 years of age, height increases by 5 centimetres and weight increases by 2.3 kilograms per year until the growth spurt during adolescence.

Psychosocial development can be measured using the Denver Developmental Screening Test. This test evaluates motor, language, and social development skills in children up to 6 years of age.

Adolescent growth

The adolescent growth spurt is closely associated with the development of the reproductive system. Puberty occurs in American girls starting at 10 or 11 years of age (average) and in American boys at age 11. In girls the first sign of puberty is the breast bud followed by breast and pubic hair development. In boys it is growth of the testes with reddening and wrinkling of the scrotum. Pubic hair appears within six months of these first signs of puberty, followed in another six months or so by enlargement of the penis.

Hearing is evaluated early, and a disorder should be suspected if speech is delayed or abnormal. Vision testing is begun in the newborn to detect strabismus and other congenital abnormalities. Visual acuity can be evaluated in children 2 to 3 years of age. Dental appointments should begin when the child is 2 or 3, because the eruption of primary teeth is usually complete by 2 years of age. Permanent teeth begin erupting about age 6 and are all in place by age 12 or 13 years.

Geriatric. The number of people in the United States older than 65 years of age is increasing rapidly, and demographers project that soon 50 percent of the American population will live to 85 years or older. As the body ages there is a steady loss in organ reserve (ability to function beyond the level normally required, which may be called upon in an emergency), which leads to decreasing functional capacity and increasing vulnerability to disease and disability. Age-related changes include the following:

1. Cellular changes occur, including decreased function and number.
2. Increased collagen results in greater stiffness and decreased tissue elasticity.
3. Muscle mass decreases, as does the mass of the liver, brain, and kidneys.
4. Cardiac output is reduced, the ability to respond to stress diminishes, and blood flow to the kidneys and other organs decreases.
5. Pulmonary function decreases because the number of alveoli lessens, expiratory muscles weaken, and there is a reduction in elastic recoil.
6. Gastrointestinal changes occur, including decreased secretion of stomach acid; decreased intestinal motil-

ity, resulting in constipation and dehydration of the stools; slower metabolism of drugs by the liver; increased incidence of gallstones; and loss of teeth, impairing proper chewing and digestion. Diverticulosis occurs in more than 50 percent of persons by age 80.

7. Excretory function diminishes because of a decrease in kidney mass and in the number of functioning nephrons.
8. Endocrine changes are noted and can include decreased functioning of thyroid and adrenal glands and decreased insulin production by the pancreas along with increasing insulin resistance that results in diabetes mellitus.
9. Neurological changes occur, including a slowing of nerve conduction velocity, a loss of brain substance, a reduction in the amount of deep sleep and an increase in the number of brief arousals, and a decrease in cerebral blood flow.
10. Visual acuity, hearing, taste, and smell decline. Vision is much more limited in dim light. The incidence of glaucoma and cataracts increases.
11. Height decreases because of the narrowing of the intervertebral disks and narrowing of the vertebrae, resulting in the loss of five centimetres by the age of 70 years.

Osteoporosis, which is demineralization of bone and loss of bone mass, results in an increased risk of fracture, especially of the hip, wrist, and spine. Bone loss is accelerated in women during menopause but can be prevented by administration of estrogen and calcium. Progesterone is added to prevent endometrial cancer if the uterus is still present.

Cancers occur most frequently in the elderly. Carcinoma of the colon is predominantly a disease of the geriatric population and is the second leading cause of death from cancer in the United States.

Depression and other mood disorders are more common among older individuals than among younger persons. The symptoms of depression may be more vague and are more likely to occur as physical symptoms than in other age groups.

Dementia (loss of intellectual function) is common among the elderly, and Alzheimer's disease is thought to account for more than 60 percent of these cases. Alzheimer's disease is characterized by a slowly progressive cognitive decline, in the absence of other causes of dementia. It affects about 10 percent of all persons older than 65 years of age.

Tests and diagnostic procedures

CLINICAL LABORATORY TESTS

Laboratory tests can be valuable aids in making a diagnosis, but, as screening tools for detecting hidden disease in asymptomatic individuals, their usefulness is limited.

Sensitivity and specificity of tests

The value of a test as a diagnostic aid depends on its sensitivity and specificity. Sensitivity is the measure of the percentage of individuals with the disease who have a positive test result (*i.e.,* people with the disease who are correctly identified by the procedure), and specificity is the measure of the percentage of people without the disease who have a negative test result (*i.e.,* healthy individuals correctly identified as free of the disease). If a test is 100 percent sensitive and the test result is negative, it can be said with certainty that the person does not have the disease, because there will be no false-negative results. If the test is not specific enough, however, it will yield a large number of false-positive results (positive test results for those who do not have the disease). The ideal test would be 100 percent sensitive and 100 percent specific; an example would be an early pregnancy test that was so accurate that it was positive in every woman who was pregnant and was never positive in a woman who was not pregnant. Unfortunately no such test exists. The normal value for a test is based on 95 percent of the population tested being free of disease, meaning that 1 out of every 20 test results in healthy individuals will be outside the normal range and therefore positive for the disease.

In the past, physicians would order selected tests based on the likelihood that the person had a certain disease. With the advent of automated analyzers, an increasing number and variety of tests have been made available at greatly reduced cost so that as many as 18 or more tests can be performed for what it previously cost to carry out three or four individual tests. A panel of chemical tests for blood and urine have become routine components of the basic medical workup. A disadvantage of this strategy is that each test produces some false-positive results and requires additional tests to rule out these diseases. The trend is reversing to perform only those tests most likely to be cost-effective.

A normal laboratory value is one that falls within a range that represents most healthy individuals. It is clear, however, that some healthy persons will have values outside that range and some individuals with disease will have values within the normal range. Thus, no sharp line divides normal and abnormal values. Tables of normal reference values must be updated regularly to react to changes in laboratory technique. Many normal values vary dramatically with age and gender.

SI units

Worldwide, the standard for reporting laboratory measurements is the International System of Units (SI units). The United States is the only major industrialized country that has not adopted the International System and continues to use customary units of measurement. Most tables provide both units to facilitate communication and understanding.

Body fluid tests. *Blood.* Blood is composed of plasma and blood cells. The blood cells—erythrocytes (red blood cells), leukocytes (white blood cells), and thrombocytes (platelets)—are suspended in the plasma with other particulate matter. Plasma is a clear, yellowish fluid that makes up more than half the volume of blood. It is distinguished from serum, which is the clear, cell-free fluid from which fibrinogen has been removed. Tests to measure the concentration of substances in the blood may use plasma, serum, or whole blood that has been anticoagulated to keep all the contents in suspension.

Complete blood count

A complete blood count (CBC) is a measure of the hematologic parameters of the blood (see the table for reference values). Included in the CBC is the calculation of the number of red blood cells (red blood cell count) or white blood cells (white blood cell count) in a cubic millimetre (mm^3) of blood, a differential white blood cell count, a hemoglobin assay, a hematocrit, calculations of red cell volume, and a platelet count.

The differential white blood cell count includes measurements of the different types of white blood cells that constitute the total white blood cell count: the band neutrophils, segmented neutrophils, lymphocytes, monocytes, eosinophils, and basophils. A specific infection can be suspected based on the type of leukocyte that has an abnor-

Reference Values in Hematology*

component	SI units	conventional units
Red blood cell count		
Female	$4.2–5.4 \times 10^{12}$/l	$4.2–5.4 \times 10^{3}$/mm^3
Male	$4.6–6.2 \times 10^{12}$/l	$4.6–6.2 \times 10^{3}$/mm^3
White blood cell count	$4.5–11.0 \times 10^{9}$/l	4,500–11,000/mm^3
Differential white blood cell count		
Band neutrophils	150–400/mm^3	3%–5%
Segmented neutrophils	3,000–5,800/mm^3	54%–62%
Lymphocytes	1,500–3,000/mm^3	25%–33%
Monocytes	300–500/mm^3	3%–7%
Eosinophils	50–250/mm^3	1%–3%
Basophils	15–50/mm^3	0%–1%
Hemoglobin		
Female	120–160 g/l	12.0–16.0 g/dl
Male	130–180 g/l	13.0–18.0 g/dl
Hematocrit		
Female	0.37–0.47	37%–47%
Male	0.40–0.54	40%–54%
Mean corpuscular volume	80–96 femtolitres	80–96 μm^3
Reticulocyte count	$25–75 \times 10^{9}$/l	25,000–75,000/mm^3
Platelet count	$150–350 \times 10^{9}$/l	$150–350 \times 10^{3}$/mm^3
Prothrombin time	12–14 seconds	12–14 seconds
Partial thromboplastin time	20–35 seconds	20–35 seconds
Plasma fibrinogen	2.0–4.0 g/l	200–400 mg/dl
Erythrocyte sedimentation rate		
Female	0–20 mm/h	0–20 mm/h
Male	0–15 mm/h	0–15 mm/h

*All values given for adults.

mal value. Viral infections usually affect the lymphocyte count, whereas bacterial infections increase the percentage of band neutrophils. Eosinophils are increased in those with allergic conditions and parasitic infection. Infection with the human immunodeficiency virus (HIV), which causes acquired immunodeficiency syndrome (AIDS), damages the body's ability to fight infection. The immune system of a healthy individual responds to infection by increasing the number of white blood cells, while the immune system of a person infected with HIV is unable to mount a defense of white blood cells (namely, lymphocytes) and cannot defend the body against viral or bacterial assault.

Of the calculations of red cell volume, the mean corpuscular volume (MCV) is the most useful for indicating anemia. The reticulocyte count, which measures the number of young red blood cells being produced, is used to distinguish between anemias resulting from a decrease in production of erythrocytes and those caused by an increase in destruction or loss of erythrocytes. An increase in the number of red blood cells (polycythemia) is normal for persons living at high altitudes, but in most of the population it indicates disease.

Platelets, small structures that are two to four micrometres in diameter, play a role in blood clotting. A decrease in the platelet count can result in bleeding if the number falls to a value below 50×10^3 per cubic millimetre.

Hematopoiesis (the production of blood cells) occurs in the bone marrow, and many types of blood disorders can be diagnosed best by analyzing a sample of bone marrow removed by a needle from the centre of the pelvic bone or the sternum (bone marrow biopsy).

Bleeding disorders are suspected when blood is seen in the skin (purpura) or a wound is delayed in clotting. In addition to a low platelet count in the peripheral blood, there may be a decrease in megakaryocytes, cells in the bone marrow that form platelets. A bleeding time greater than 20 minutes indicates an abnormality of platelet function. Other screening tests for coagulation disorders include the prothrombin time (PT) test, the partial thromboplastin time (PTT) test, and the plasma fibrinogen assay (see the table). Blood factors, which are protein elements essential to the clotting of blood, should be assayed if a disorder associated with one of them is suspected. For example, factor VIII or IX can be assayed if the patient is thought to have hemophilia.

The erythrocyte sedimentation rate (ESR) is the rate at which red blood cells settle in a column of blood in one hour. It is a nonspecific indicator of inflammatory disease that is also increased in anemia (see the table).

The Coombs, or antiglobulin, test (AGT) is used to test blood cells for compatibility when doing a cross match to detect antibodies that would interfere with a blood transfusion. It also is used to detect antibodies to red blood cells in hemolytic disease of the newborn and drug-induced hemolytic anemias.

Urine. Examining the urine is one of the oldest forms of diagnostic testing, extending back to the days of Hippocrates. Physicians observed the urine (uroscopy) to diagnose all forms of illness because direct examination of a patient, or at least disrobing the patient, was socially unacceptable (see above *Historical aspects*).

Urinalysis

Urinalysis is the most commonly performed test in the physician's office. It consists of (1) a gross examination, in which the colour, turbidity, and specific gravity of the urine are assessed, (2) the use of a dipstick (a plastic strip containing reagent pads) to test for bilirubin, blood, glucose, ketones, leukocyte esterase, nitrite, pH, protein, and urobilinogen, and (3) a microscopic examination of a centrifuged specimen to detect red or white blood cells, casts, crystals, and bacteria. The urine is collected using a "clean-catch" technique to eliminate contamination with bacteria from skin or vaginal secretions.

Dipstick tests are available that contain from 2 to 10 different tests. The test for glucose, which indicates diabetes, and the test for protein, which indicates kidney disease, tumours of the urinary tract, or hypertensive disorders of pregnancy, are two of the most important assays available.

The microscopic examination is the most valuable urinalysis test. It will show a variety of cells that are normally shed from the urinary tract. Usually up to five white blood cells per high-power field (HPF) are present; more than 10 white blood cells per HPF indicates a urinary tract infection. More than two red blood cells per HPF is abnormal, although in women this is often due to vaginal contamination from menstruation. Cylindrically shaped urinary casts, shed from the kidney's tubules, consist of protein mixed with cells or other materials and may indicate renal disease if present in large numbers. Various crystals also are found in the urinary sediment, but these are generally of little clinical significance.

Feces. The tests most commonly performed on feces are the fecal occult blood test (FOBT), stool cultures, and the examination for parasites. The fecal occult blood test is a low-cost method for detecting bleeding, which may be the first sign of carcinoma of the colon or rectum. Although the false-positive rate for this test is low, the false-negative rate is high. It is more likely to detect lesions in the right (ascending) colon because they bleed more than those in the left (descending) colon. Routine surveillance for colorectal cancer depends on periodic fecal occult blood testing combined with direct visualization of the lower colon with a sigmoidoscope (see below).

Fecal occult blood test

Individuals who are at increased risk for colon cancer and should be screened regularly are identified by any of the following criteria: age greater than 50 years, previous colorectal cancer or adenoma, family history of colon cancer or polyps in a first-degree relative or another genetic predisposition (*e.g.*, cancer family syndrome), history of ulcerative colitis or Crohn's disease, or personal or family history of genital or breast cancer.

Stool cultures are obtained when diarrhea is severe and particular bacteria such as *Salmonella, Shigella,* or *Giardia* are suspected. If a parasitic infection is suspected, the stool is examined under the microscope for the eggs or cysts of parasites such as pinworms (*Enterobius vermicularis*) or roundworms (*Ascaris lumbricoides*).

Cerebrospinal fluid. Examination of the cerebrospinal fluid, obtained by lumbar puncture (*i.e.,* a needle inserted into the lower back), is performed if meningitis or hemorrhage into the central nervous system (subarachnoid hemorrhage) is suspected. The fluid is normally crystal clear and colourless. It will contain blood if subarachnoid hemorrhage has occurred.

Tests give clues to various disease processes. Viral meningitis can be differentiated from bacterial meningitis by the type of white blood cells identified, although a bacterial culture is the definitive test. The glucose value will usually be normal in patients with viral meningitis but low in those with bacterial and fungal meningitis. The protein level is increased in individuals with meningitis and tumour. The pressure of the fluid within the spinal canal is measured after the needle is inserted. The pressure is elevated in the presence of infection and tumour.

Gastric fluid. By passing a tube through the nose and into the stomach, gastric fluid can be obtained from the stomach for examination. The most common reason for this test is to look for blood in the upper gastrointestinal tract. Gastric fluid also can be cultured to test for tuberculosis if an adequate sputum sample cannot be obtained for culture.

Semen. More than 10 percent of couples in the United States have difficulty establishing a pregnancy. In addition to obtaining a complete history, performing a physical examination of both partners, and verifying that ovulation does occur in the woman, the physician will perform a semen analysis. Normal semen contains more than 60 million sperm per millilitre. More than 60 percent of the sperm are motile two hours after ejaculation, and 80 to 90 percent will have normal form and structure. Possible causes of infertility are a low sperm count, low motility, or a low percentage of normal forms.

Infertility in men

Miscellaneous tests. *Immunologic procedures.* Immunologic blood tests demonstrate abnormalities of the immune system. Immunity to disease depends on the body's ability to produce antibodies (immunoglobulins) when challenged by foreign substances (antigens). Antibodies bind to and help eliminate antigens from the body.

The inability of the body to produce certain classes of immunoglobulins (IgG, IgA, IgM, IgD, IgE) can lead to disease. Complexes formed by the antigen-antibody reaction can be deposited in almost any tissue and can lead to malfunction of that organ. Immunofluorescence assays to detect antinuclear antibodies (antibodies that will bind to antigens within the nucleus) can be used to diagnose systemic lupus erythematosus and rheumatoid arthritis.

The inability of the body to develop antibodies to invading bacteria may result from infection with HIV, which invades white blood cells—primarily monocytes, macrophages, and helper T lymphocytes. Helper T cells are a subgroup of T lymphocytes that are the primary regulators of the immune response and proliferate in response to antigenic stimulation. Testing for HIV is performed with an enzyme-linked immunosorbent assay (ELISA) and the western immunoblotting antibody test (western blot).

Diabetes mellitus

Oral glucose tolerance test. The glucose tolerance test is used to confirm or exclude the diagnosis of diabetes mellitus when a fasting blood glucose test result is not definitive (*i.e.,* greater than the upper range of the normal value, 115 milligrams per decilitre [mg/dl; 6.4 mmol/l], but less than the diagnostic level for diabetes, 140 mg/dl [7.8 mmol/l]). Even if a blood glucose test is obtained after fasting 10 to 12 hours and the level is above 140 mg/dl, it is important to confirm the result with a second determination to rule out other factors that may have given a one-time abnormal test result.

The oral glucose tolerance test measures the response of the body to a challenge load (an amount calculated to evoke a response) of glucose. It most often is used during pregnancy to detect early glucose intolerance that could pose a significant risk to the infant. After a fasting blood glucose test result is obtained, 75 grams of glucose (100 grams if the patient is pregnant) is administered and blood samples are taken every 30 minutes for two hours. In patients with diabetes, the blood glucose value will rise to a higher level and remain higher longer than in individuals who do not have diabetes.

A simpler but less reliable screening test is the two-hour postprandial blood glucose test, performed two hours after intake of a standard glucose solution or a meal containing 100 grams of carbohydrates. A plasma glucose level above 140 mg/dl indicates the need for a glucose tolerance test.

Gastrointestinal absorption tests. Malabsorption of nutrients can result from surgical alterations or physiological disturbances of the gastrointestinal tract: removal of a significant portion of the bowel can cause the malabsorption condition short-bowel syndrome, a diffuse mucosal disease such as sprue can interfere with absorption, and diseases of the liver or pancreas may prevent digestive enzymes from reaching the intestines. Bacterial overgrowth in the intestines can interfere with glucose absorption, and the stomach's failure to produce intrinsic factor will prevent the absorption of vitamin B_{12} (cobalamin), which leads to pernicious anemia.

Persons who have a low serum vitamin B_{12} level and who are suspected of having pernicious anemia usually are required to undergo the Schilling test. Radioactive vitamin B_{12} is administered orally, and the amount excreted in the urine over the next 24 hours is measured. Malabsorption is confirmed if less than 8 percent of the vitamin B_{12} is excreted in the urine.

Steatorrhea is the excretion of an excessive amount of fat in the stool, which is diagnostic of fat malabsorption when the amount of fat in the diet is normal. Stool specimens are collected for three days following two days of a diet containing 100 grams of fat per day. The excretion of more than six grams of fat daily indicates fat malabsorption, which may occur in persons with pancreatic disease, in those with diffuse mucosal disease, and in those who have undergone massive small-bowel resection.

A five-carbon sugar, D-xylose, is absorbed in the duodenum and proximal jejunum. It is not metabolized and is excreted unchanged in the urine. The D-xylose absorption test measures the absorption ability of the jejunum. Lowered excretion indicates diminished intestinal absorption usually caused by a decreased absorptive surface, infiltrative intestinal disease, or bacterial overgrowth.

Toxicological tests. Toxicology is the study of poisons—their action, detection, and the treatment of conditions they produce. Many substances are toxic only at high concentrations. Lithium, for example, is used to treat bipolar (manic-depressive) disorder but can be toxic at high levels. Another example is acetaminophen, which is valuable in controlling fever and discomfort but is toxic in large doses.

Levels of toxicity

The concentration of an element in the blood is the usual measure of toxicity. The therapeutic blood range is the concentration of the drug that provides therapeutic benefit, whereas the toxic blood range is the concentration at which toxic manifestations are likely.

Some substances such as insecticides are toxic to one individual and not to another. Many environmental substances as well as some encountered in the workplace are toxic in high doses; these include organic solvents, heavy metals, mineral dusts, dyes, and cigarette smoke. Acceptable exposure levels are controlled by government standards.

The nervous system is most sensitive to toxicological damage. Common toxins that cause damage to peripheral nerves are the six-carbon solvents, such as *n*-hexane, in glues or solvents and organophosphorus compounds. Carbon disulfide, used in the production of rayon fibres and cellophane, is a potent neurotoxin. Because no specific treatment is available for most of these toxic manifestations, preventing overexposure is important.

GENETIC TESTING

The diagnostic evaluation of a genetic disorder begins with a medical history, a physical examination, and the construction of a family pedigree documenting the diseases and genetic disorders present in the past three generations. This pedigree aids in determining if the problem is sex-linked, dominant, recessive, or not likely to be genetic.

Chemical, radiological, histopathologic, and electrodiagnostic procedures can diagnose basic defects in patients suspected of genetic disease. These include chromosome karyotyping (in which chromosomes are arranged according to a standard classification scheme), enzyme or hormone assays, amino acid chromatography of blood and urine, gene and deoxyribonucleic acid (DNA) probes, blood and Rh typing, immunoglobulin determination, electrodiagnostics, and hemoglobin electrophoresis.

As a result of genetic mutation, a genetic disorder can occur in a child with parents who are not affected by this disorder. This mutation can occur when the egg or sperm form (germinal mutation), or it can occur later following conception, when chromosomes from the egg and sperm combine. Mutations can occur spontaneously or be stimulated by such environmental factors as radiation or carcinogenic (cancer-causing) agents. Mutations occur with increasing frequency as people age. In men this may result from errors that occur throughout a lifetime as DNA replicates to produce sperm. In women nondisjunction of chromosomes becomes more common later in life, increasing the risk of aneuploidy (too many or too few chromosomes). Long-term exposure to ambient ionizing radiation may cause genetic mutations in either gender.

Cytogenetics

Cytogenetics is the microscopic study of chromosomes and the transmission of genetic material from parent to offspring. Humans have 22 pairs of identical chromosomes plus a pair of sex chromosomes (one inherited from each parent). There are 30,000 to 40,000 genes arranged along the chromosomes in linear order, each having a precise location, or locus. The goal of the international human genome project was to map the location of all genes by the year 2005; a rough map was produced in 2000 and a finished map in 2003. The mapping of specific locations of variations in the human genome will help to identify the genetic causes of a number of diseases.

Two broad classes of genes have been implicated in the development of cancer—oncogenes, which promote tumour growth, and tumour-suppressor genes. Both types of cancer-related genes, usually more than one variation of each type, are involved in a particular cancer, such as that of the colon or breast.

Prenatal diagnosis. Prenatal screening is performed if there is a family history of inherited disease, the mother is at an advanced age, a previous child had a chromosomal

abnormality, or there is an ethnic indication of risk (Ashkenazic Jews and French Canadians are at increased risk for Tay-Sachs disease; blacks, Arabs, Turks, and others for sickle-cell anemia; and those of Mediterranean descent for thalassemia [hereditary anemia]). Parents can be tested before or after conception to determine whether they are carriers.

Alpha-fetoprotein screening

The most common screening test is an assay of alpha-fetoprotein (AFP) in the maternal serum. Elevated levels are associated with neural tube defects in the fetus such as spina bifida (defective closure of the spine) and anencephaly (absence of brain tissue). When alpha-fetoprotein levels are elevated, a more specific diagnosis is attempted using ultrasonography and amniocentesis to analyze the amniotic fluid for the presence of alpha-fetoprotein and acetylcholinesterase. Fetal cells contained in the amniotic fluid also can be cultured and the karyotype (chromosome morphology) determined to identify chromosomal abnormality. Cells for chromosome analysis also can be obtained by chorionic villus sampling, the direct needle aspiration of cells from the chorionic villus (future placenta). (See REPRODUCTION AND REPRODUCTIVE SYSTEMS: *Human reproduction from conception to birth. The normal events of pregnancy: Prenatal care and testing.*)

Chromosomal analysis. To obtain a person's karyotype, laboratory technicians grow human cells in tissue culture media. After being stained and sorted, the chromosomes are counted and displayed. The cells are obtained from the blood, skin, or bone marrow or by amniocentesis or chorionic villus sampling, as noted above. The standard karyotype shown in the procedure has approximately 400 visible bands, and each band contains up to several hundred genes.

When a chromosomal aberration is identified, it allows for a more accurate prediction of the risk of its recurrence in future offspring. Karyotyping can be used not only to diagnose aneuploidy, which is responsible for Down, Turner's, and Klinefelter's syndromes, but also to identify the chromosomal aberrations associated with solid tumours such as Wilms' tumour, meningioma, neuroblastoma, retinoblastoma, renal-cell carcinoma, small-cell lung cancer, and certain leukemias and lymphomas.

DNA probes. Karyotyping requires a great deal of time and effort and may not always provide conclusive information. It is most useful in identifying very large defects involving hundreds or even thousands of genes.

Newer techniques such as fluorescent in situ hybridization (FISH) have much higher rates of sensitivity and specificity. FISH also provides results more quickly because no cell culture is required. This technique can detect smaller genetic deletions involving one to five genes. It is also useful in detecting moderate-sized deletions such as those causing Prader-Willi syndrome, which is characterized by a rounded face, low forehead, and mental retardation.

Gene analysis with recombinant DNA

The analysis of individual genes has been greatly enhanced by the development of recombinant DNA technology. Small DNA fragments can be isolated, and unlimited amounts of cloned material can be produced. Once cloned, the various genes and gene products can be used to study gene function in healthy individuals and those with disease. Recombinant DNA methods can detect any change in DNA, down to a one-base-pair change out of the three billion base pairs in the genome. DNA probes are labeled with radioactive isotopes or fluorescent dyes and used to identify persons who are carriers for autosomal recessive conditions. Disorders that can be detected using this technique include hemophilia A, polycystic kidney disease, sickle-cell disease, Huntington's chorea, cystic fibrosis, and hemochromatosis.

Biochemical tests. Biochemical tests primarily detect enzymatic defects such as phenylketonuria, porphyria, and glycogen-storage disease. Although testing of newborns for all these abnormalities is possible, it is not cost-effective, because some are quite rare. Screening requirements for these disorders vary from state to state and depend on whether the disease is sufficiently common, has severe consequences, and can be treated or prevented if diagnosed early and whether the test can be applied to the entire population at risk.

INSTRUMENTAL SCREENING

Scopes. *Sigmoidoscopy.* Colorectal cancer is the second leading cause of death from cancer in the United States. This disease is preventable if adenomatous polyps, protruding growths from the mucosal surface that can progress to cancer, are identified and removed. Although most adenomatous polyps are not cancerous, this possibility can only be discounted by histologic examination, which requires their removal. Fifty percent of all lesions occur in the rectum and sigmoid colon; they can be detected and removed using a 60-centimetre flexible sigmoidoscope. This instrument consists of a bundle of optical fibres that carry the visual image; it can be bent at the tip in four directions using controls at the base so that it can be maneuvered through the contorted sigmoid colon. The scope also contains a light source at the tip for illuminating the bowel, as well as separate passageways for instilling air and water, for suctioning fluid, and for inserting such instruments as biopsy forceps and snares. This scope has a smaller diameter than do rigid scopes and causes the patient less discomfort because of its flexibility. The operator can see the organ directly through a magnifying eyepiece or indirectly by a video monitor. The latter allows videotaping of suspicious lesions. Both rigid and flexible scopes can be fitted with a still camera.

The flexible fibre-optic sigmoidoscope comes in lengths of 35 and 60 centimetres. When fully inserted, the 60-centimetre scope can reach to the mid-descending colon and is the more frequently used scope. The colonoscope is a similar flexible fibre-optic scope that is longer and can reach the cecum, thus allowing evaluation of the entire colon. Its use requires that the patient be sedated because its passage through the entire colon is more uncomfortable.

A rigid, 25-centimetre sigmoidoscope is less expensive and allows direct visualization of the bowel, but it is less popular because of the greater discomfort its rigidness causes. The proctoscope and anoscope, shorter rigid instruments used to visualize the lower rectum and anus, are used to diagnose and treat hemorrhoids and other lesions in the anorectal area.

Screening for colon cancer

The incidence of colon cancer increases sharply after the age of 50. Asymptomatic individuals should have a sigmoidoscopy at age 50 and, if the result is negative, the test should be repeated every three to five years. Symptomatic persons and those with a family history of colon cancer should start regular examinations at age 40 or younger.

Esophagogastroduodenoscopy. As the lengthy name implies, esophagogastroduodenoscopy (EGD) is an endoscopic examination in which a scope is passed through the esophagus, stomach, and duodenum for a visual examination. This flexible fibre-optic scope contains the same channels as the flexible fibre-optic sigmoidoscope described above and usually has a camera attached to record visually recognizable abnormalities.

This procedure is indicated when symptoms of peptic ulcer disease persist despite an adequate trial of treatment or when there is upper gastrointestinal bleeding or a suspicion of upper gastrointestinal cancer. It is also indicated if there is an esophageal stricture or obstruction or persistent vomiting of unknown cause. Esophageal strictures, if benign, can be successfully dilated, and upper gastrointestinal bleeding can be controlled using electrocoagulation. If the bleeding is from esophageal varices, they can be injected with a sclerosing (hardening) agent. A tissue sample can be removed and examined (a biopsy) from any suspicious lesion of the esophagus, stomach, or duodenum to make the specific tissue diagnosis that is necessary when deciding on the most appropriate therapy.

Endoscopic retrograde cholangiopancreatoscopy. The flexible fibre-optic scope used in endoscopic retrograde cholangiopancreatography (ERCP) is similar to the scopes described above. It is passed through the stomach into the duodenum to visualize the ampulla of Vater, the opening of the common bile duct into the duodenum. It enables injection of a radiopaque dye into the common bile duct to permit radiographic visualization of the common bile duct and the pancreatic duct. This test is used to evaluate the patient with jaundice whose biliary tract is suspected to be obstructed because of a gallstone or tumour. It is also used

to evaluate persistent pancreatitis of unknown cause. If there is stricture of the ampulla or another area in the common bile duct, a sphincterotomy (incision of the sphincter) or balloon dilatation can be used to enlarge the opening.

Laparoscopy. Fibre-optic technology has greatly expanded the procedures that can be performed by laparoscopy. By using local anesthesia and mild sedation, the abdominal wall can be punctured and the laparoscope inserted to examine the contents of the abdomen, obviating the need for major surgery and general anesthesia. Instruments are inserted through multiple ports in the abdomen, and surgeons can visualize abdominal organs without making an open incision into the abdomen. Valuable diagnostic information can be obtained by examining a biopsy specimen of the liver or abdominal lesions. Surgeons also can perform a variety of procedures with this method, such as removing the gallbladder and ligating the fallopian tubes. In orthopedic surgery the same technique is called arthroscopy, and it simplifies the treatment of many disorders that previously required a large surgical incision and a lengthy period of rehabilitation.

Alternative to major surgery

Nasopharyngolaryngoscopy. The use of fibre-optic nasopharyngolaryngoscopes permits visualization of structures inside the nasal passages such as the sinus openings, larynx, and vocal cords. A more thorough examination can be performed than is possible using indirect visualization with a mirror.

Colposcopy. The colposcope is a lighted magnifying scope used to directly visualize the vulva, vagina, and cervix and to evaluate suspicious areas. Colposcopy is used when the Papanicolaou test suggests the possibility of cancer; it helps to detect precancerous abnormalities and identifies in which areas a biopsy should be performed for a definitive diagnosis to be made.

Graphing and miscellaneous instrumental screening. *Electroencephalogram.* The electroencephalogram (EEG) is a record of electrical activity of the brain recorded by 8 to 16 pairs of electrodes attached to the scalp. It is useful in the diagnosis of epilepsy, brain tumours, and sleep disorders and in the assessment of patients with suspected brain death. The latter use is particularly important if organs are to be saved for transplantation as soon as brain death is confirmed. Sleep deprivation and other provocative tests, including photic stimulation and hyperventilation, can be used to accentuate borderline findings. The EEG is of no use in diagnosing psychiatric illness.

Electrocardiogram. The electrocardiogram (ECG) is a graphic recording of the electrical activity of the heart detected at the body surface and amplified. It was invented by the Dutch physiologist Willem Einthoven (1860–1927) and for many years was called an EKG after the German *Elektrokardiogramm*. Electrodes to record the electrical activity of the heart are placed at 10 different locations, one on each of the four limbs and six at different locations on the anterior chest wall. Twelve different leads, or electrical pictures, are generated, each having its own normal configuration.

Evaluating the heart

The ECG is of greatest use in diagnosing cardiac arrhythmias, acute and prior myocardial infarctions, pericardial disease, cardiac enlargement (atrial and ventricular), and various electrolyte disturbances and drug effects. The exercise electrocardiogram, or ECG stress test, is used to assess the ability of the coronary arteries to deliver oxygen while the heart is undergoing strain imposed by a standardized exercise protocol. If the blood supply to the heart is jeopardized during exercise, the inadequate oxygenation of the heart muscle is recorded by typical changes in the electrocardiogram that indicate coronary artery disease (narrowing of the coronary arteries).

Echocardiography. The echocardiogram is a noninvasive technique used to record the structure of the heart by using ultrahigh-frequency sound waves. A transducer placed on the chest wall emits a short burst of ultrasound waves and then measures the reflection, or echo of the sound as it bounces back from such cardiac structures as valves and the muscle wall. It is used to evaluate chamber size, wall thickness, wall motion, valve structure, and valve motion. It is the method of choice for detecting infection of the valves (endocarditis), intracardiac tumours, and pericardial fluid. Mitral valve prolapse is easily visualized by this noninvasive technique.

Myocardial perfusion imaging. Myocardial perfusion imaging uses radioactive thallium to detect myocardial ischemia, myocardial infarction, and coronary artery disease. Injected intravenously, radioactive thallium is rapidly absorbed by the myocardium and is normally distributed evenly in heart muscle. Deficient blood flow to a portion of the myocardium is readily detectable by decreased uptake in that area. Evidence of recent and not-so-recent myocardial infarcts will be visible, but most persons with coronary artery disease who have not had a previous infarction will have normal perfusion patterns when they are at rest. In such a patient a thallium stress test is performed in which the substance is injected while the individual is exercising so that areas of transient ischemia can be identified and the patient treated to prevent myocardial infarction. An alternative means of stressing the heart that can provide information comparable with exercise is the injection of dipyridamole, a vasodilator. This test is used to diagnose coronary artery disease when the resting electrocardiogram is abnormal or the exercise electrocardiogram is equivocal.

Another method for evaluating the heart without the stress of exercise involves the intravenous injection of the drug dobutamine, while monitoring the effects using echocardiography. By using dobutamine echocardiography, the heart condition of frail patients and those who have heart disease or physical limitations that preclude exercise can be evaluated. Dobutamine induces the same changes in the heart that would occur during a standard exercise test. Two-dimensional echocardiography shows areas of the left ventricle that function abnormally. This technique uses no X-ray or radioactive material and is useful in diagnosing heart disease during pregnancy (see above *Echocardiography*).

Cardiac catheterization and angiography. A more specific measurement of coronary artery narrowing is carried out by placing a catheter into the heart though which a radiopaque dye is injected, allowing the cardiac chambers and coronary arteries to be directly visualized. This test is more expensive and more hazardous than the noninvasive procedures and is usually performed after the others to quantify the severity of disease present and to establish whether the person is a candidate for surgical intervention with balloon angioplasty or coronary bypass surgery. It is also used to evaluate patients with suspected valvular disease and those with angina who do not respond to treatment.

Electromyography. Electromyography (EMG), the graphing and study of the electrical characteristics of muscles, is used to differentiate disease of the muscles from disease of the peripheral nerves. A needle electrode is inserted into the muscle, and the electrical activity of the muscle is measured. Resting muscle is normally electrically silent. The electrical potential is measured with the muscle at rest and during contraction. The response to electrical stimulation allows the physician to determine whether muscle weakness is the result of a disease in the muscle, such as a myositis (inflammation of the muscle), or a disease of nerves leading to muscle (neuropathy), such as Guillain-Barré syndrome.

Diagnosing muscle disorders

SURGICAL EXAMINATION

Biopsy. A biopsy is the removal of tissue for microscopic examination to establish a precise diagnosis. Tissue can be obtained from any organ by excision, incision, removal by a needle, or scraping. Glass slides of the tissue are prepared and examined microscopically to define the characteristic nature of the lesion.

An excisional biopsy is the total removal of the lesion to be examined and is most often used to diagnose skin lesions. The major advantage of excisional biopsy is that it provides the pathologist with the entire lesion and minimizes the chance that a cancer in part of the lesion would be missed. This technique is practical only when the lesion is accessible and is less than two or three centimetres in diameter.

An incisional biopsy involves the removal of only a por-

tion of the lesion for pathological examination and is used when the size or location of the tumour prohibits its complete excision. This technique also is used when a needle biopsy does not provide adequate information for a diagnosis to be made.

A needle biopsy is the simplest and least disruptive way to obtain tissue for pathological examination. This procedure can be performed using either a large cutting needle to obtain a "core" of tissue or a small-gauge needle. The latter technique, termed fine-needle aspiration biopsy, is accomplished by inserting the needle into the area of interest and applying suction to draw the tissue into the needle. A needle biopsy is often used to obtain specimens from breast masses. It is less expensive and involves less morbidity than does an open biopsy. The main disadvantages include the missing of deep lesions with the needle and the need for a specially trained pathologist to accurately interpret the specimen. As noted above, often more cells are needed for a precise diagnosis than are provided by a fine-needle biopsy.

Fine-needle aspiration biopsy

Another form of aspiration biopsy is the endometrial biopsy, in which the specimen is obtained by applying suction through a curette inserted into the uterus to obtain cells from the internal lining.

Abrasion is a method by which cells are obtained from the surface of lesions using a brush or spatula. Cells from epithelial-lined body cavities and surfaces such as the vagina, bronchus, and stomach are examined using the Papanicolaou technique. The Papanicolaou test or smear, commonly called the Pap smear, is the examination of cervical cells that have been fixed and stained on a slide according to the technique developed by the Greek physician George Nicolas Papanicolaou. This technique also can be applied to cells obtained from other surfaces.

Exploratory surgery. When a specific diagnosis is not possible using noninvasive or simple biopsy techniques, it may be necessary to surgically explore the area in question. If the lesion is in the abdomen, this involves a laparotomy or incision into the abdomen to observe the lesion. If possible a biopsy sample is removed. It may be apparent that the lesion is inoperable because of its location or attachment to vital structures from which it cannot be separated.

RADIOLOGICAL SCREENING

Named after Wilhelm Conrad Röntgen, the roentgenogram is the photograph of internal structures made by passing X rays through the body to produce a shadow image on specially sensitized film. The value of a roentgenogram is considerably enhanced by the use of contrast material, such as barium, to make structures visible on the film that would otherwise not appear. Perhaps the most common procedure employs a barium enema, administered to the patient before the X-ray examination, which allows identification of polyps as small as one centimetre in diameter when air is inserted after the barium (a double-contrast barium enema). This screening is effective if precancerous polyps are identified at an early stage.

Chest film. One of the most common screening roentgenograms is the chest film, taken to look for such infections as tuberculosis and conditions like heart disease and cancer. Treatment of tuberculosis detected by a roentgenogram can prevent more extensive infection, but unfortunately this technique is of little value in screening for lung cancer because the stage at which the disease is detectable by this method is too far advanced for treatment to be of value.

Screening for breast cancer

Mammography. New film screening techniques make it possible to detect lesions in the breast using low doses of radiation. Mammography is never a substitute for a clinical breast examination by a physician, because not all lesions are detectable by X-ray examination; however, lesions often can be detected by mammography before they are palpable in the breast. The primary purpose for mammography is the detection of cancer at the earliest, treatable stage, before the lesion is palpable.

Mammography is most useful in older women whose breast tissue is less dense than that of younger women. Mammography is never a substitute for a biopsy if a suspicious mass is palpated. Some groups recommend an initial mammogram at 35 to 40 years of age to serve as a baseline for subsequent screening. The American Cancer Society recommends a mammogram every one to two years from age 40 to 49 and yearly thereafter. However, women at increased risk for breast cancer should consider initiating annual mammographic screening before the age of 40. The risk of breast cancer is doubled or trebled in women who have a sister with breast cancer or whose mother was diagnosed with breast cancer before the age of 40.

COMPUTERIZED BODY SCANNING

Computed tomography. The introduction of computed tomography (CT scan) in 1972 was a major advance in visualizing almost all parts of the body. Particularly useful in diagnosing tumours and other space-occupying lesions, it uses a tiny X-ray beam that traverses the body in an axial plane. Detectors record the strength of the exiting X rays; this information is then processed by a computer and a cross-sectional image of the body produced.

CT is the preferred examination for evaluating stroke, particularly subarachnoid hemorrhage, as well as abdominal tumours and abscesses.

Ultrasonography. Ultrasonography, or ultrasound imaging, uses pulsed or continuous high-frequency sound waves to image internal structures by recording the differing reflection signals. The sonographic image is not as precise as images obtained through computed tomography or magnetic resonance imaging, but it is used in many procedures because it is quick and relatively inexpensive and has no known biological hazards when used within the diagnostic range.

Advantages of ultrasound

This method is used to diagnose gallstones, heart defects, and tumours. It is used to guide certain procedures such as needle biopsies and the introduction of tubes for drainage. It has become an essential part of obstetric and prenatal assessment, although controversy exists as to its routine use in obstetric care. Ultrasonography plays an integral role in the diagnosis and management of fetal abnormalities; it is also used to guide intrauterine corrective surgery.

Magnetic resonance imaging. Magnetic resonance imaging (MRI) relies on the response of magnetic fields to short bursts of radio-frequency waves to produce computer images that provide structural and biochemical information about tissue. The process uses radio waves and is thus much safer than imaging using X rays or gamma rays. This totally noninvasive but very expensive procedure is particularly useful in detecting cerebral edema, abnormalities of the spine, and early-stage cancer. In examining the brain, spinal cord, urinary bladder, pelvic organs, and cancellous bone, MRI is the superior imaging technique. Because patients must lie quietly inside a narrow tube, MRI may raise anxiety levels in the patients, especially those with claustrophobia. Another disadvantage of MRI is that it has a longer scanning time than CT, which makes it more sensitive to motion artifacts and thus of less value in scanning the chest or abdomen. Because of the strong magnetic field, MRI cannot be used if a pacemaker is present or if metal is present in critical areas such as the eye or brain.

MRI has largely supplanted arthrography, the injection of dye into a joint to visualize cartilage or ligament damage to the knee or shoulder, and myelography, the injection of dye into the spinal canal to visualize spinal cord or intervertebral disk abnormalities.

Multiple sclerosis, a disease with multiple foci of demyelination (loss of the myelin sheath of a nerve) in the brain, sometimes can be diagnosed using MRI. However, because the test is not sufficiently sensitive, a normal MRI cannot exclude the diagnosis.

Magnetic resonance angiography, a unique form of MRI technology, can be used to produce an image of flowing blood. This permits the visualization of arteries and veins without the need for needles, catheters, or contrast agents.

CT and MRI provide two-dimensional views of cross sections of the body, and these images must be viewed in sequence by the radiologist. Computer technology now makes it possible to construct holograms that provide three-dimensional images from digital data obtained by conventional CT or MRI scanners. These holograms can

Three-dimensional images

be useful in locating lesions more precisely and in mapping the exact location of coronary arteries when planning bypass surgery or angioplasty.

Digital subtraction angiography. Digital subtraction angiography (DSA), an electronic technique for imaging blood vessels, is useful in diagnosing arterial occlusion, including carotid artery stenosis and pulmonary artery thrombosis, and in detecting renal vascular disease. After contrast material is injected into an artery or vein, a physician produces fluoroscopic images. Using these digitized images, a computer subtracts the image made with contrast material from a postinjection image made without contrast material, producing an image that allows the dye in the arteries to be seen more clearly. In this manner, the images arising from soft tissues, bones, and gas are the same in the initial and subsequent image and are thereby eliminated by the subtraction process. The remaining images of blood vessels containing the contrast material are thus more prominent.

Positron emission tomography. Positron emission tomography (PET) is a highly sensitive technique for diagnosing stroke and other neurological diseases such as multiple sclerosis and epilepsy. Positron-emitting radionuclides with short half-lives are used to detect cerebral blood flow, oxygen utilization, and glucose metabolism, providing both qualitative and quantitative information regarding metabolism and blood flow, such as in the heart.

Formulating a diagnosis

The process of formulating a diagnosis is called clinical decision making. The clinician uses the information gathered from the history and physical examination to develop a list of possible causes of the disorder, called the differential diagnosis. The clinician then decides what tests to order to help refine the list or identify the specific disease responsible for the patient's complaints. During this process, some possible diseases (hypotheses) will be discarded and new ones added as tests either confirm or deny the possibility that a given disease is present. The list is refined until the physician feels justified in moving forward to treatment. Even after treatment is begun, the list of possible diagnoses may be revised further if the patient does not progress as expected.

The differential diagnosis

The hypotheses are ranked with the most likely disease placed first. Sometimes, however, a less likely disease is addressed first because it is more life-threatening and could lead to serious consequences if not treated promptly. Following this course, the possibility of a heart attack would be eliminated first in a patient experiencing chest pain and appendicitis would be the first condition to be addressed in a child with abdominal pain, even though another less serious disease is more likely.

An algorithm is a sequence of alternate steps that can be taken to solve problems—a decision tree. Starting with a chief complaint or key clue, the physician moves along this decision tree, directed one of two ways by each new piece of information, and eliminates diagnoses. If the wrong path is taken, the physician returns to a previous branching point and follows the other path. Computers can be used to assist in making the diagnosis; however, they lack the intuition of an experienced physician and the nonverbal diagnostic clues obtained during the interview.

Diagnostic tests rarely establish the presence of a disease without doubt. The greater the sensitivity and the specificity of the test, the more useful it will be. Ordering too many tests poses significant danger, not only because of low cost-effectiveness but also because a falsely abnormal test result requires a further series of tests to prove or disprove its accuracy. This further testing may involve additional discomfort, risk, and cost to the patient, which is especially unfortunate if the tests need not have been ordered in the first place. It is just as important to know when not to order a test as to know which tests to order.

An important feature of clinical decision making is the ongoing relationship between the physician and patient. The knowledge a physician gains in caring for the patient for a long period of time can provide greater insight into the likelihood of a given disease being present. When the symptoms are caused by emotional factors, the familiar personal physician is more likely to accurately diagnose them than is a physician seeing the patient for the first time. Also, a lengthy and trusting association with a physician will often positively influence the patient's outcome. Thus, sporadic visits to the emergency department of a hospital, where physicians who are unfamiliar with the patient are asked to provide diagnoses and treatment without the benefit of this partnership, are more likely to be inefficient, expensive, and less personally satisfying.

Physician-patient relationship

Early in the course of a disease, decisions must be made with fewer clues to the diagnosis than are likely to be available later. One of the most difficult tasks in medicine is to separate, in the early stages of an illness, the serious and life-threatening diseases from the transient and minor ones. Many illnesses will resolve without a diagnosis ever being reached. Nevertheless, an illness may remain undiagnosed for months or years before new symptoms appear and the disease advances to a stage that permits diagnosis.

Patients often have undifferentiated complaints that can represent an uncommon serious disorder or a common but not very serious disorder. For example, a patient may experience fatigue. Depending on the patient's family history and personal background, the physician may think initially of depression and next of anemia secondary to gastrointestinal bleeding. A variety of less likely disorders will follow. Anemia is easy to rule out with inexpensive hemoglobin and hematocrit tests. These tests should be ordered even if depression is the correct diagnosis because anemia may contribute to the weariness and should be treated as well. Depression can be diagnosed with appropriate questioning, and a physical examination may eliminate many other diagnostic possibilities.

THERAPEUTICS

Preventive medicine

The rationale for preventive medicine is to identify risk factors in each individual and reduce or eliminate those risks in an attempt to prevent disease. Primary prevention is the preemptive behavior that seeks to avert disease before it develops—for example, vaccinating children against diseases. Secondary prevention is the early detection of disease or its precursors before symptoms appear, with the aim of preventing or curing it. Examples include regular cervical Papanicolaou test screening and mammography. Tertiary prevention is an attempt to stop or limit the spread of disease that is already present. Clearly, primary prevention is the most cost-effective method of controlling disease.

The five leading causes of death in the United States are cardiovascular disease, cancer, cerebrovascular disease, accidental injuries, and chronic lung disease. The single most preventable cause of death in the United States is cigarette smoking, which is linked to cardiovascular disease (heart attack), cancer (lung, larynx, bladder, pancreas, and so on), cerebrovascular disease (stroke), and chronic lung disease (emphysema, chronic bronchitis).

Leading causes of death in the United States

Following earlier work by the Canadian Task Force on the Periodic Health Examination, the U.S. Preventive Services Task Force was established to evaluate the effectiveness of various screening tests, immunizations, and prophylactic regimens based on a critical review of the scientific literature. Its report, *Guide to Clinical Preventive Services,* lists the recommendations for the 60 target conditions evaluated by the panel.

Immunization is the best method for preventing infectious diseases. Standard immunizations of infants and children include those for diphtheria, tetanus, and pertussis (DTP); polio (OPV); measles, mumps, and rubella (MMR); *Haemophilus influenzae* type b (HbCV); and hep-

atitis B (HBV). A yearly vaccine against the influenza virus should be administered to adults who are older than 65 years of age, to those at risk because of chronic cardiopulmonary disease, and to those in chronic care facilities. Adults also should be immunized once at age 65 years against pneumococcal pneumonia with a vaccine containing 23 of the most common strains of *Streptococcus pneumoniae*.

Acquired immunodeficiency syndrome (AIDS), caused by the human immunodeficiency virus (HIV), is also a major infectious disease problem. Although a vaccine is expected, obstacles to its development are great. The only primary preventive measures currently available are either to abstain from sexual contact or to use condoms and, among intravenous drug users, to avoid sharing needles. Almost 25 percent of adult AIDS cases in the United States are related to infection from needles used to administer illegal drugs.

Preventing the spread of AIDS

The risk factors for coronary artery disease that can be modified to prevent myocardial infarction are cigarette smoking, hypertension, an elevated serum cholesterol level, a sedentary lifestyle, obesity, stress, and excessive alcohol consumption. In addition to an elevated total serum cholesterol level, an elevated low-density lipoprotein (LDL) level and a decreased high-density lipoprotein (HDL) level are significant risk factors. The total cholesterol level and elevated LDL level can be reduced by appropriate diet, whereas a low HDL can be raised by stopping smoking and increasing activity. If these measures do not provide adequate control, a variety of drugs capable of lowering the cholesterol level are available.

The major risk factor for stroke is hypertension, with cigarette smoking and diabetes mellitus significantly increasing the risk. Transient ischemic attacks (TIAs) occur before stroke in 20 percent of patients and consist of sudden onset of one or more of the following symptoms: temporary loss of vision in one eye, unilateral numbness, temporary loss of speech or slurred speech, and localized weakness of an arm or leg. Attacks last less than 24 hours and resolve without permanent damage until the stroke occurs.

Relation of smoking and cancer

The most important preventive behaviour in averting cancer is the avoidance of cigarette smoke. Smoking accounts for 30 percent of all cancer deaths, and there is increasing recognition of the danger of environmental or sidestream smoke to the nonsmoker. Primary prevention of skin cancer includes restricting exposure to ultraviolet light by using sunscreens or protective clothing. Secondary preventive measures include mammography, clinical breast examinations, and breast self-examinations for breast cancer; pelvic examinations and Papanicolaou tests for cervical and ovarian cancer; and sigmoidoscopy, digital rectal examinations, and stool tests for occult blood for colorectal cancer.

Demineralization of bone and a reduction in bone mass (osteoporosis) occur most often in men and women age 70 or older and may result in fractures, low back pain, and loss of stature. Osteoporosis in postmenopausal women that is caused by estrogen deficiency is the most common manifestation. The most effective method for preventing loss of bone mass after menopause is estrogen replacement therapy and increased calcium intake. Primary preventive measures include increasing physical activity and avoiding cigarettes and heavy alcohol consumption.

Alcohol abuse is the primary reason that accidents are the fourth leading cause of death in the United States. Other factors are failure to wear seat belts or motorcycle helmets, sleep deprivation, and guns in the home. Taking reasonable precautions and being aware of the potential dangers of alcohol and firearms can help reduce the number of deaths due to accidents.

Treatment of symptoms

PAIN

Pain is the most common of all symptoms and often requires treatment before its specific cause is known. Pain is both an emotional and a physical experience and is difficult to compare from one person to another. One patient may have a high pain threshold and complain only after the disease process has progressed beyond its early stage, while another with a low pain threshold may complain about pain that would be ignored or tolerated by most people. Pain from any cause can be increased by anxiety, fear, depression, loneliness, and frustration or anger.

Acute and chronic pain

Acute pain serves a useful function as a protective mechanism that leads to the removal of the source of the pain, whether it be localized injury or infection. Chronic pain serves a less useful function and is often more difficult to treat. Although acute pain requires immediate attention, its cause is usually easily found, whereas chronic pain complaints may be more vague and difficult to isolate.

The ideal method for treating pain is to eliminate the cause, such as to surgically remove an inflamed structure, to apply hot compresses to a muscle spasm, or to set a fractured bone in a cast. Alternatives to drug therapy, such as physical therapy, should be relied on whenever possible. The analgesic drugs most often used to alleviate mild and moderate pain are the nonsteroidal anti-inflammatory drugs (NSAIDs) such as aspirin, ibuprofen, acetaminophen, or indomethacin. If these are ineffective, a weak opiate such as codeine, hydrocodone, or oxycodone would be the next choice. Severe pain not controlled by these agents requires a strong opiate such as morphine or meperidine. Because opiates are addictive, their use is controlled by the Controlled Substances Act, and individuals prescribing or dispensing these drugs must register annually with the Drug Enforcement Administration. Each drug is assigned to one of five groups, from schedule I, which includes drugs that have the highest potential for abuse, to schedule V, which includes drugs with a limited dependence-causing potential.

NAUSEA AND VOMITING

Nausea and vomiting are common symptoms that may arise from diseases of the gastrointestinal tract, including gastroenteritis or bowel obstruction, from medications, such as analgesics or digoxin, or from nervous system disturbances such as migraine headaches or motion sickness. Vomiting is controlled by a vomiting centre located in the medulla oblongata of the brain stem.

Identifying and treating the cause is important, especially if the condition responds well to treatment and is serious if not addressed. A bowel obstruction can occur as a result of adhesions from previous abdominal surgery. Obstruction or decreased bowel motility also can occur with constipation and fecal impaction. Such important and treatable causes must be ruled out before resorting to antiemetic (serving to prevent or cure vomiting) drugs. The most frequently used antiemetic agents are the phenothiazines, the most popular being prochlorperazine (Compazine [trademark]). Antihistamines may be useful in motion sickness, but newer and more powerful drugs are needed to control the vomiting associated with cancer chemotherapy.

Nausea and vomiting are experienced by more than 50 percent of pregnant women during the first trimester. These symptoms are referred to as morning sickness, although they can occur at any time of the day. They may be distressing, but they cause no adverse effect on the fetus. Drug therapy is not only unnecessary; it should be avoided unless proved safe for the fetus. Treatment involves rest and intake of frequent small meals and pyridoxine (vitamin B_6).

Treatment of morning sickness

DIARRHEA

Acute diarrhea can result from food poisoning, laxatives, alcohol, and some antacids but usually is caused by an acute infection with bacteria such as *Escherichia coli, Salmonella,* and *Staphylococcus aureus.* In infants, acute diarrhea is usually self-limiting, and treatment consists primarily of preventing dehydration. Traveler's diarrhea affects up to half of those traveling to developing areas of the world. Preventive measures include chewing two tablets of bismuth subsalicylate (Pepto-Bismol [trademark]) four times a day, drinking only bottled water or other bottled or canned beverages, and eating only those fruits that have been peeled, canned products, and restaurant food that is piping hot. Avoiding dairy products, raw seafood

and vegetables, and food served at room temperature also limits exposure. Severe cases require antibiotic therapy.

COUGH

Coughing is a normal reflex that helps clear the respiratory tract of secretions and foreign material. It also can result from irritation of the airway or from stimulation of receptors in the lung, diaphragm, ear (tympanic membrane), and stomach. The most common cause of acute cough is the common cold. Chronic cough is most often caused by irritation and excessive mucus production that results from cigarette smoking or from postnasal drainage associated with an allergic reaction.

Treatment includes humidification of the air to loosen secretions and to counteract the drying effect of coughing and inflammation. Moist air from a vaporizer or a hot shower helps, as do hot drinks and soups. Antihistamines are often used to treat acute cough, but their value is questionable if an allergy is not present. They may also cause additional drying of the respiratory mucosa. Guaifenesin is widely used in cough preparations to help liquefy secretions and aid expectoration. Decongestants reduce secretions by causing vasoconstriction of the nasopharyngeal mucosa. The most common decongestants found in many cough preparations are pseudoephedrine, phenylephrine, and phenylpropanolamine. They may cause high blood pressure, restlessness, and urinary retention and should be used with caution in anyone being treated for hypertension. Narcotics are powerful cough suppressants, codeine being the most frequently used. Several safer nonnarcotic antitussive (cough-preventing) agents are available such as dextromethorphan, which has almost equal effectiveness but fewer side effects. Most cough preparations containing dextromethorphan also contain a decongestant and an expectorant. Because coughing is an important defense mechanism in clearing secretions from blocked airways, a productive cough (one that produces secretions) should not be suppressed.

Side effects of cough preparations

INSOMNIA

Insomnia is a difficulty in falling asleep or the feeling that sleep is not refreshing. Transient insomnia occurs when there are stressful life events or schedule changes, as shift workers or those who travel across multiple time zones experience. A disturbed sleep can also be related to the intake of stimulating drugs, anxiety, depression, or medical conditions associated with pain. Anxiety usually causes difficulty in falling asleep, whereas depression is associated with early morning awakening. The elderly spend less time sleeping, and their sleep is lighter and marked by more frequent awakenings. This situation is exacerbated by afternoon napping.

The treatment of insomnia involves establishing good sleep hygiene: maintaining a consistent schedule of when to retire and awaken, setting a comfortable room temperature, and minimizing such disruptive stimuli as noise and light. Daily exercise is beneficial but should be avoided immediately before bedtime. Stimulants should be avoided, including nicotine and caffeine. Alcohol disrupts the normal sleep pattern and should also be avoided. Drinkers sleep more lightly and frequently awaken unknowingly, which leaves them feeling unrefreshed the next day.

When medication is required, physicians usually prescribe one of the sleep-inducing benzodiazepines. They may have long-, intermediate-, or ultrashort-acting effects. None should be used regularly for long periods. Various nonbenzodiazepine hypnotics and sedatives are also available, and their usefulness varies according to individual preference.

Designing a therapeutic regimen

Once the physician makes a diagnosis or identifies the most likely cause of the symptoms and decides on the appropriate treatment, an entirely new set of conditions becomes operative. One of the first conditions to be considered is the patient's reason for seeking medical advice and the patient's expectations. The patient's visit may have been precipitated by the discovery that a friend's minor symptom, similar in nature to one the patient has been experiencing, proved to be something serious. If tests can rule out this possibility, reassurance may serve as a therapeutic action. When possible, physicians work to cure a disease and thereby relieve the symptoms, but many times the disease is unknown or chronic and incurable. In either case, relief from or improvement of symptoms or restoration of normal functioning is the goal. When neither a cure nor complete relief of symptoms is possible, an explanation of the disease and knowledge of the cause and what to expect may provide significant relief. Patients often want to know the name of the disease, what caused it, how long it will last, what additional symptoms may occur, and what they can do to assist the physician's treatment to hasten recovery. Providing information about the disease can help to alleviate anxiety and fears that could otherwise impede the patient's progress.

Attention to the patient's emotional well-being

An essential ingredient of any successful therapeutic regimen is the positive attitude of the patient toward the physician. A relationship of trust and respect for the physician based on reputation or years of supportive care is one of the physician's most powerful therapeutic tools.

When selecting a management plan, the physician usually has several options, and the outcomes or consequences of each will vary. Often, the best choice is one made together with the patient, who may have definite preferences for a trial of therapy over further testing or for oral medication rather than an injection, even if the latter would provide more rapid relief. The possible side effects of the medicine or treatment may well influence therapeutic choice, such as if a person would prefer dizziness to nausea. Once a course of therapy is selected, a new decision tree arises that leads to new options, depending on the response. Further testing, increasing the dose of medication, or changing to a new drug may be required. Almost every treatment has some degree of risk, from either unwanted side effects or unexpected complications. The physician describes these risks in terms of probability, expecting the patient to accept or reject the treatment based on these odds and his or her willingness to suffer the side effects or to risk the complications to achieve relief.

Another factor affecting therapeutic success is patient compliance—the degree to which patients adhere to the regimen recommended by their physician. Therapeutic regimens that require significant changes in lifestyle, such as recommendations to follow a special diet, begin an exercise program, or discontinue harmful habits like smoking cigarettes, are likely to result in poor compliance. Also, the greater the number of drugs prescribed and the more complicated the regimen, the poorer is the compliance. A patient is much more likely to successfully follow a regimen of taking a single dose of medication daily than one prescribed four times daily. Patients also may not fully realize the need to continue taking the medication after their symptoms have subsided, despite a physician's instruction to finish the medicine. Patient compliance may be most difficult to achieve in chronic but generally asymptomatic illnesses such as hypertension. Patients who experience no symptoms may need to be convinced of the necessity of taking their medication daily to prevent the occurrence of an untoward event (in hypertension, a stroke or other cardiovascular problems). Similarly, patients with depression or anxiety may want to discontinue medication once their symptoms abate. Until a relapse occurs, they may not recognize the need to continue taking the medication until instructed to taper the dosage slowly.

Adhering to the therapeutic regimen

In deciding which therapeutic regimen is likely to be most effective, the physician must depend on scientific studies that compare one drug or treatment regimen with others that have been proved effective. The most dependable study is one that is truly objective and removes the possibility of bias on the part of the patient who wants the drug to work and the bias of the physician who may expect a certain outcome and subtly influence the interpretation. Such a study is "double-blind"; it controls for both possible tendencies by comparing an active drug with an inactive "look-alike" drug. Neither the patient nor the physician knows which drug the patient is taking, so that neither one's bias can influence the result. Although this is

The double-blind study

the best way to demonstrate the effectiveness of a drug, it is sometimes very difficult to control for all the variables that could influence the outcome, such as varying degrees of stress one group or another may be under. Physicians will use the results of a wide variety of studies similar to this study to decide whether a regimen or drug is likely to work in a given patient; however, they will depend most heavily on their past experience with drugs or other techniques that have worked under similar circumstances. It is knowledge based on experience and on understanding of the patient that leads to the greatest therapeutic success.

Diet

PROPHYLACTIC MEASURES OF NUTRITION

General requirements. Adequate nutritional intake is required to maintain health and prevent disease. Certain nutrients are essential; without them a deficiency disease will result. Required nutrients that cannot be synthesized by the body and therefore must be taken regularly are essential amino acids, water-soluble and fat-soluble vitamins, minerals, and essential fatty acids. The U.S. Recommended Dietary Allowances (RDAs), one of many sets of recommendations put out by various countries and organizations, have been established for these essential nutrients by the Food and Nutrition Board of the National Academy of Sciences. These RDAs are guidelines and not absolute minimums. Intake of less than the RDA for a given nutrient increases the risk of inadequate intake and a deficiency disorder. Nutritional requirements are greater during the periods of rapid growth (infancy, childhood, and adolescence) and during pregnancy and lactation. Requirements vary with physical activity, aging, infections, medications, metabolic disorders (*e.g.*, hyperthyroidism), and other medical situations. RDAs do not address all circumstances and are designed only for the average healthy person.

Protein, needed to maintain body function and structure, consists of nine essential amino acids that must be provided from different foods in a mixed diet. Ten to 15 percent of calories should come from protein. The oxidation of 1 gram (0.036 ounce) of protein provides 4 kilocalories of energy. The same is true for carbohydrate, but fat yields 9 kilocalories.

Carbohydrate provides about 45 percent of calories in the American diet, in the form of sugars, starches (complex carbohydrates), and dietary fibre (indigestible carbohydrates). Fibre is not digestible but increases the bulk of the stool and facilitates faster intestinal transit, which some believe reduces the risk of colon cancer by diminishing the time that cancer-producing substances in the diet remain in contact with the bowel wall. Increasing bulk also decreases the concentration of these substances. Dietary fibre can be insoluble (wheat bran) or soluble (oat bran and psyllium). Only the soluble fibres found in oats, fruit, and legumes lower blood cholesterol and benefit individuals with diabetes by delaying the absorption of glucose.

The most concentrated source of energy is fat, the source of fat-soluble vitamins and essential fatty acids. Thirty-seven percent of calories in the American diet come from fat, but the ideal is closer to 30 percent. The average American diet also contains 450 milligrams daily of cholesterol, but less than 300 milligrams is recommended.

Food Guide Pyramid

The recommended daily diet as determined by the U.S. Department of Agriculture is called the Food Guide Pyramid and consists of 6 to 11 servings of bread, cereal, rice, or pasta; 3 to 5 servings of vegetables; 2 to 4 servings of fruit; 2 to 3 servings of fish, meat, poultry, dry beans, eggs, or nuts; and 2 to 3 servings of milk, yogurt, or cheese.

Requirements in infancy. Nutritional needs are greatest during the first year of life. Meeting the energy demands during this period of rapid growth requires 100 to 120 kilocalories per kilogram per day. Breast milk, the ideal food, is not only readily available at the proper temperature, it also contains antibodies from the mother that help protect against disease. Infant formulas closely approximate the contents of breast milk, and both contain about 50 percent of calories from carbohydrate, 40 percent from fat, and 10 percent from protein.

Breast milk or commercial formula is recommended for the first six months of life and may be continued through the first year. Solid foods are introduced at four to six months of age starting with rice cereal and then introducing a new vegetable, fruit, or meat each week. Cow's milk should not be given to infants younger than one year of age, and low-fat milk should be avoided throughout infancy because it does not contain adequate calories and polyunsaturated fats required for development. Additional iron and vitamins should be given, especially to infants at high risk of iron deficiency, such as those with a low birth weight.

Toddlers are usually picky eaters, but attempts should be made to include the following four basic food groups in their diet: meat, fish, poultry, or eggs; dairy products such as milk or cheese; fruits and vegetables; and cereals, rice, or potatoes. Mealtime presents an excellent opportunity for social interaction and strengthening of the family unit. This starts with the bonding between mother and child during breast-feeding and continues as a source of family interaction throughout childhood.

Requirements in adolescence. Nutritional needs during adolescence vary according to activity level, with some athletes requiring an extremely high-calorie diet. Other adolescents, however, who are relatively sedentary consume calories in excess of their energy needs and become obese. Peer pressure and the desire for social acceptance can profoundly affect the quality of nutrition of the adolescent as food intake may shift from the home to fast-food establishments.

Risks of adolescent pregnancy

Pregnancy during adolescence can present special hazards if the pregnancy occurs before the adolescent has finished growing and if she has established poor eating habits. Pregnancy increases the already high requirements for calcium, iron, and vitamins in these teenagers.

Eating disorders such as anorexia nervosa and bulimia arise predominantly in young women as a result of biological, psychological, and social factors. An excessive concern with body image and a fear of becoming fat are hallmarks of these conditions. The patient with anorexia nervosa has a distorted body image and an inordinate fear of gaining weight; consequently she reduces her nutritional intake below the amount needed to maintain a normal minimal weight. Severe electrolyte disturbances and death can result. Bulimia is a behavioral disorder marked by binge eating followed by acts of purging (*e.g.*, self-induced vomiting, ingestion of laxatives or diuretics, or vigorous exercising) to avoid weight gain.

Requirements of the elderly. The elderly often have decreased intestinal motility and decreased gastric acid secretion that can lead to nutritional deficiencies. The problem can be accentuated by poorly fitting dentures, poor appetite, and a decreased sense of taste and smell. Although lower levels of activity reduce the need for calories, older persons may feel something is wrong if they do not have the appetite of their younger years, even if caloric intake is adequate to maintain weight. The reduction in gastric acid secretion can lead to decreased absorption of vitamins and other nutrients. Nutritional deficiencies can reduce the level of cognitive functioning. Vitamin supplementation, especially with cobalamin (vitamin B_{12}), may be particularly valuable in the elderly.

The diet of the geriatric population is often deficient in calcium and iron, with the average woman ingesting only half the amount of calcium needed daily. Decreased intake of vegetables can also contribute to various nutritional deficiencies.

Constipation, which is common in the elderly, results from decreased intestinal motility and immobility and is worsened by reduced fluid and fibre intake. The multiple medications that the elderly are likely to be taking may contribute to constipation and prevent the absorption of certain nutrients. Some drugs, such as the phenothiazines, may interfere with temperature regulation and lead to problems during hot weather, especially if fluid intake is inadequate.

Requirements in pregnancy. The growing fetus depends on the mother for all nutrition and increases the mother's usual demand for certain substances such as iron, folic

acid, and calcium, which should be added as supplements to a balanced diet that contains most of the other required nutrients. The diet of adolescent girls, however, is often deficient in calcium, iron, and vitamins. If poor nutritional habits have been established previously and are maintained during pregnancy, the pregnant adolescent and her fetus are at increased risk.

In addition to avoiding junk foods, the pregnant woman should abstain from alcohol, smoking, and illicit drugs because these all have a detrimental effect on the fetus. Although the average recommended weight gain during pregnancy is approximately 11.3 kilograms (25 pounds), the pregnant woman should be less concerned with a maximum weight gain than she is with meeting the nutritional requirements of pregnancy. Low weight gain (less than 9.1 kilograms) has been associated with intrauterine growth retardation and prematurity in the United States.

Requirements during breast-feeding

Women who are breast-feeding should continue taking vitamin supplements and increasing their intake of calcium and protein to provide adequate breast milk. This regimen will not interfere with the mother's ability to slowly lose the weight gained during pregnancy.

THERAPEUTIC MEASURES OF NUTRITION

Changes in diet can have a therapeutic effect on obesity, diabetes mellitus, hypertension, peptic ulcer, and osteoporosis.

Obesity. About one-fourth of the American population meets the definition of obesity (20 percent above ideal body weight). Obesity occurs when the number of calories consumed exceeds the number that is metabolized, the remainder being stored as adipose (fat) tissue. Many theories address the causes of obesity, but no single cause is apparent. Multiple factors influence weight, including genetic factors, endocrine levels, activity levels, metabolic rates, eating patterns, and stress.

The treatment of obesity requires reducing calorie intake while increasing calorie expenditure (exercise). Because obesity is a chronic illness, it requires long-term lifestyle changes unless surgery is performed to effect permanent changes in the digestion of food. Thus fad diets, no matter how effective they are in the short term, remain inadequate for long-term weight control. A reduction in calorie intake of 500 kilocalories per day should lead to a loss of 0.45 kilogram (1 pound) per week. This reduction can be increased by greater calorie reduction or an accompanying exercise program. With exercise, the weight loss will be primarily fat, whereas without it, muscle is lost as well. Exercise also leads to a "positive" addiction that makes it easier to sustain regular exercising for long periods. It reduces the risk of heart disease and can improve self-esteem.

Weight-reduction diets for the obese individual should be similar to those used by nonobese persons but with fewer calories—namely, a low-fat diet that avoids high-calorie foods. One of the most popular and successful of these diets is the very-low-calorie diet (VLCD) that results in rapid fat loss while minimizing the loss of lean muscle tissue. These diets require supplementation with potassium and a vitamin-mineral complex. Fad diets that eliminate one foodstuff, such as carbohydrate or protein, may give short-term results but fail in the long term to maintain the weight loss. Furthermore, these diets can lead to medically significant problems, such as ketosis (a buildup of ketones in the body).

Very-low-calorie diet

Appetite-suppressing drugs have limited short-term and no long-term effectiveness. Surgery can provide long-term benefits but it is an option only to those at least 45.3 kilograms heavier than their ideal body weight who are willing to suffer the common complications. The most frequently performed procedures are vertical banded gastroplasty and gastric bypass, both of which effectively reduce the size of the stomach.

Diabetes mellitus. Diet is the cornerstone of diabetic treatment whether or not insulin is prescribed. The goal is to regulate the patient's blood glucose level to as close to normal as possible and for the patient to achieve and maintain an ideal weight. Refined and simple sugars are avoided, and saturated fat is reduced by focusing the diet on poultry and fish rather than meat as a major source of protein. Soluble fibre such as that found in beans and oatmeal is recommended in contrast to the insoluble fibre found in wheat and bran. Artificial sweeteners are effective low-calorie replacements for simple sugar. The American Diabetes Association's recommendations are similar to those of the American Heart Association—that is, adhering to a balanced diet with restricted saturated fat intake while maintaining normal weight. Three or four meals of equal caloric content are spaced throughout the day, especially when supplemental insulin is needed.

Hypertension. Many patients with hypertension benefit from a low-sodium diet (reduced sodium chloride [table salt] intake), and physicians often recommend this as part of the initial therapy for hypertension. If alterations in diet fail to counteract the hypertension, drugs such as diuretics may be prescribed along with potassium supplements (because diuretics may deplete potassium). Other dietary measures are directed toward achieving an ideal body weight because obesity contributes to hypertension and increases the risk of cardiovascular disease. An adequate low-sodium diet can be achieved with a no-added-salt diet—that is, no salt is added to food after preparation, and foods with a high-sodium content such as cured meats are avoided. Low-sodium diets should be combined with increased potassium, which can be obtained by eating fruits, especially bananas, and vegetables or using salt substitutes.

Low-sodium diet

Peptic ulcer. In the past a bland diet and frequent ingestion of milk and cream were the mainstays of ulcer treatment. Today the only dietary regimen is the avoidance of such irritating foods as spicy and highly seasoned foods and coffee. The newer drug therapies decrease gastric acidity much more than antacids and other dietary measures do. The infection of the stomach by *Helicobacter pylori* is now recognzied as a major factor in chronic gastritis and recurrent peptic ulcer in many patients. This bacterial infection requires a treatment regimen consisting of antibiotics and a bismuth-containing compound, which is different from the treatment of an ulcer that is not caused by *H. pylori*.

Osteoporosis. Although little can be done to treat osteoporosis once it is established, a great deal can be accomplished to prevent it, as has been discussed above (see above *Preventive medicine*). Osteoporosis, the loss of bone density, occurs in men and women older than 70 years of age and is manifested primarily in hip, wrist, and vertebral fractures. It is most noticeable in postmenopausal women who have not taken estrogen. Estrogen replacement therapy, which should be combined with supplemental calcium, is most effective in decreasing bone resorption when begun during menopause, although it will provide some benefit if started later. In women who have an intact uterus, estrogen must be taken with progesterone to reduce the risk of endometrial cancer.

Biological therapy

BLOOD AND BLOOD CELLS

Blood transfusions were not clinically useful until about 1900 when the blood types A, B, and O were identified and cross-matching of the donor's blood against that of the recipient to prove compatibility became possible. When blood with the A antigen (type A or AB) is given to someone with anti-A antibodies (type B or O blood), lysis of the red blood cells occurs, which can be fatal. Persons with blood type O are universal donors because this blood type does not contain antigen A or B; however, because type O blood contains antibodies against both A and B, patients with this blood type can receive only type O blood. Fortunately, type O is the most common blood type, occurring in 40 to 60 percent of people, depending on the selected population (*e.g.,* 40 percent of the white population has blood type O, while 60 percent of Native Americans have it). Conversely, persons with type AB blood are universal recipients. Having no antibodies against A or B, they can receive type O, A, or B red blood cells.

Blood transfusions

Most individuals are Rh-positive, which means they have the D antigen; less than 15 percent of the population lack this antigen and are Rh-negative. Although anti-D antibodies are not naturally present, the antigen is so highly

immunogenic (able to provoke an immune response) that anti-D antibodies will develop if an Rh-negative person is transfused with Rh-positive blood. Severe lysis of red blood cells will occur at any subsequent transfusion. The condition erythroblastosis fetalis, or hemolytic disease of the newborn, occurs when Rh-positive babies are born to Rh-negative mothers who have developed anti-D antibodies either from a previous transfusion or by maternal-fetal exchange during a previous pregnancy.

Whole blood transfusions are infrequently used because most transfusions only require one or more specific blood components. Whole blood, which contains red blood cells, plasma, platelets, and coagulation factors, is used mainly during cardiac surgery and when there is moderate or massive hemorrhage. It can be used only up to 35 days after it has been drawn and is not always available, because most units of collected blood are used for obtaining components.

Packed red blood cells are what remains of whole blood after the plasma and platelets have been removed. A 450-milliliter unit of whole blood is reduced to a 220-milliliter volume. Packed red blood cells are used most often to raise a low hemoglobin or hematocrit level in patients with chronic anemia or mild hemorrhage.

Leukocyte-poor red blood cells are obtained by employing a filter to remove white blood cells (leukocytes) from a unit of packed red blood cells. This type of transfusion is used to prevent febrile reactions in patients who have had multiple febrile transfusion reactions in the past, presumably to white blood cell antigens.

Platelet transfusions are used to prevent bleeding in patients with very low platelet counts, usually less than 20,000 cells per microlitre, and in those undergoing surgery whose counts are less than 50,000 cells per microlitre.

Autologous transfusion is the reinfusion of one's own blood. The blood is obtained before surgery and its use avoids transfusion reactions and transfusion-transmitted diseases. Donation can begin one month before surgery and be repeated weekly, depending on the number of units likely to be needed.

PLASMA

Coagulation factors

Plasma, the liquid portion of the blood, is more than 90 percent water. It contains all the noncellular components of whole blood including the coagulation factors, immunoglobulins, electrolytes, and proteins. When frozen, the coagulation factors remain stable for up to one year but must be used within 24 hours when thawed. Fresh frozen plasma is used in patients with multiple clotting factor deficiencies, such as in those with severe liver disease or massive hemorrhage.

Cryoprecipitate is prepared from fresh frozen plasma and contains about half the coagulation factors in $\frac{1}{15}$ the volume. It is used to treat patients with deficiencies of factor VIII, von Willebrand factor, factor XIII, and fibrinogen.

Specific clotting factor concentrates are prepared from pooled plasma or pooled cryoprecipitate. Factor VIII concentrate, the antihemophilic factor, is the preferred treatment for hemophilia A. A monoclonal antibody–purified human factor VIII is also available. Factor IX complex, the prothrombin complex, is also available for treating hemophilia B (factor IX deficiency).

IMMUNOGLOBULINS

Immune serum globulin (ISG), obtained from the plasma of unselected donors, contains a mixture of immunoglobulins, mainly IgG, with lesser amounts of IgM and IgA. It is used to provide passive immunity to a variety of diseases such as measles, hepatitis A, and hypogammaglobulinemia. Intravenous immunoglobulins (IVIGs) provide immediate antibody levels and avoid the need for painful intramuscular injections.

Hyperimmune serum globulin is prepared in the same way as the nonspecific immunoglobulin above but from patients who are selected because of their high titres of specific antibodies. Rh-immune globulin is given to pregnant Rh-negative women to prevent hemolytic disease of the newborn. Other hyperimmune serum globulins are used to prevent hepatitis B, tetanus, rabies, and varicella-zoster in exposed individuals.

BONE MARROW TRANSPLANTATION

Graft-versus-host disease

Bone marrow transplantation does not involve the transfer of a discrete anatomic organ as occurs in other forms of transplantation, but it entails the same risk of rejection by the recipient, which is called graft-versus-host disease (GVHD). The main indications for bone marrow transplantation are leukemia, aplastic anemia, and congenital immunologic defects.

Immunosuppressive drugs and irradiation are usually used to prepare the recipient. Close matching of tissue between donor and recipient is also essential to minimize GVHD, with autologous transplantation being the best method (the patients donate their own marrow at times of remission to be used later). Allogeneic (homologous) bone marrow transplants by a matched donor (preferably a sibling) are the most common.

Bone marrow transplantation is not recommended for patients older than 50 years of age, because of the higher mortality that results. The incidence of graft-versus-host disease increases in those older than 30 years of age. Those who donate bone marrow incur no risk, because they generate new marrow to replace that which has been removed. General anesthesia is required, however, to aspirate the bone marrow from the iliac crests, which is then infused into the recipient.

HEMATOPOIETIC GROWTH FACTORS

The hematopoietic growth factors are potent regulators of blood cell proliferation and development in the bone marrow. They are able to augment hematopoiesis when bone marrow dysfunction exists. Recombinant DNA technology has made it possible to clone the genes responsible for many of these factors. Some are commercially available and can be used to stimulate white blood cell development in patients with neutropenia (a decrease in the number of neutrophilic leukocytes) associated with cancer chemotherapy.

The first to be developed was erythropoietin, which stimulates red blood cell production. It is used to treat the anemia associated with chronic renal failure and that related to therapy with zidovudine (AZT) in patients infected with HIV. It may also be useful in reversing anemia in cancer patients receiving chemotherapy. Filgrastim (granulocyte colony-stimulating factor [G-CSF]) is used to stimulate the production of white blood cells, which prevents infection in patients whose white blood cells are diminished because of the effects of anticancer drugs. Another is sargramostim (granulocyte-macrophage colony-stimulating factor [GM-CSF]), which is used to increase the white blood cell count in patients with Hodgkin's disease or acute lymphoblastic leukemia who are undergoing autologous bone marrow transplantation.

BIOLOGICAL RESPONSE MODIFIERS

Biological response modifiers, used to treat cancer, exert their antitumour effects by improving host defense mechanisms against the tumour. They have a direct antiproliferative effect on tumour cells and also enhance the ability of the host to tolerate damage by toxic chemicals that may be used to destroy the cancer.

Biological response modifiers include monoclonal antibodies, immunomodulating agents such as the bacille Calmette-Guérin (BCG) vaccine used against tuberculosis, lymphokines and cytokines such as interleukin-2, and the interferons.

Role of interferon

The three major classes of interferons are interferon-α, produced by white blood cells; interferon-β, produced by fibroblasts; and interferon-γ produced by lymphocytes. The interferons are proteins produced by these cells in response to viral infections or other stimuli; they have antiviral, antiproliferative, and immunomodulatory properties that make them useful in treating some viral infections and cancers. They do not act directly on the viruses but rather indirectly, increasing the resistance of cells to viral infections. This can be particularly useful in patients who have an impaired immune system and a diminished

ability to fight viral infections, especially those with AIDS.

Interferon-α is produced by a recombinant DNA process using genetically engineered *Escherichia coli.* Recombinant interferon-α appears to be most effective against hairy-cell leukemia and chronic myelogenous leukemia, lymphoma, multiple myeloma, AIDS-associated Kaposi's sarcoma, and chronic type C hepatitis. It is moderately effective in treating melanoma, renal cell carcinoma, and carcinoid. It also can enhance the effectiveness of chemotherapy in some cancers. Unfortunately, treatment with this drug can be quite toxic.

Interferon-γ may prove useful in treating a different set of diseases—for example, chronic conditions such as rheumatoid arthritis.

HORMONES

The term hormone is derived from the Greek *hormaein,* meaning "to set in motion." It refers to a chemical substance that has a regulatory effect on a certain organ or organs. There are sex hormones such as estrogen and progesterone, thyroid hormones, insulin, adrenal cortical and pituitary hormones, and growth hormones.

Estrogens (estradiol, estone, and estriol) promote the growth and development of the female reproductive system—the vagina, uterus, fallopian tubes—and breasts. They are responsible for the development of secondary sex characteristics—growth of pubic and axillary hair, pigmentation of the nipples and genitals—and contribute to bone formation. The decrease in estrogen after menopause contributes to bone demineralization and osteoporosis, and hormone replacement therapy is often recommended to counteract this occurrence (see above *Preventive medicine*). Postmenopausal estrogen also prevents atrophic vaginitis, in which the vaginal mucosa becomes thin and friable. Estrogens can be administered orally, through the skin (transdermally), vaginally, and intramuscularly.

Progestins combined with estrogens comprise the oral contraceptives that inhibit ovulation by affecting the hypothalamus and pituitary. Progestin-only pills and injections are also effective contraceptives that work by forming a thick cervical mucus that is relatively impenetrable to sperm. Although the mortality associated with all forms of birth control is less than that associated with childbirth, this is not true for women older than the age of 35 years who smoke cigarettes. Their risk of stroke, heart attacks, and other cardiovascular problems is greatly increased, and the use of oral contraceptives is contraindicated. Levonorgestrel is a synthetic progestin that is implanted beneath the skin of the upper arm in six Silastic (trademark) capsules and provides birth control for five years.

Androgens

Androgens consist of testosterone and its derivatives, the anabolic steroids. Testosterone is produced in the testes in males, and small amounts are produced by the ovary and adrenal cortex in females. Testosterone is used to stimulate sexual organ development in androgen-deficient males and to initiate puberty in selected boys with delayed growth. The anabolic steroids are testosterone derivatives that provide anabolic activity with less stimulation of growth of the sexual organs. The use of anabolic steroids to increase muscle strength and endurance has been universally deplored by the medical community. This practice may have serious long-term effects such as the development of atherosclerotic disease because of effects on the blood lipids, especially the lowering of high-density lipoproteins. Their use in juvenile athletes can cause premature epiphyseal closure (early ossification of the growth zone of bones), compromising the attainment of their full adult height.

Human chorionic gonadotropin (HCG) is a hormone produced by cells of the placenta that can be extracted from the urine of pregnant women days after fertilization and thus is used in the early detection of pregnancy. It is also used to stimulate descent of the testicles in boys with prepubertal cryptorchidism and to treat infertility in men with underdeveloped testicles. Because it can stimulate the thyroid, it was inappropriately thought to be useful in treating obesity; there is no clinical proof of its effectiveness in this application.

Growth hormone, produced by the pituitary gland, stimulates linear growth and regulates metabolic functions. Inadequate secretion of this hormone by the pituitary will impair growth in children, which is evidenced by their poor rate of growth and delayed bone age (*i.e.,* slowed bone development). A synthetic preparation of the hormone is used to treat children who have a congenital deficiency of growth hormone.

Adrenal corticosteroids are any of the steroid hormones produced by the adrenal cortex except for the sex hormones. These include the mineralocorticoids (aldosterone) and glucocorticoids (cortisol), the secretion of which is regulated by the adrenocorticotrophic hormone (ACTH) produced in the anterior pituitary. Overproduction of ACTH leads to excessive secretion of glucocorticoids (Cushing's syndrome), which also can result from an increased concentration of corticosteroids secreted by tumours of the adrenal gland; conversely, the production of an insufficient amount of adrenal corticosteroids results in primary adrenocortical insufficiency (Addison's disease). The glucocorticoids are used primarily for their potent anti-inflammatory effects in rheumatic disorders, collagen diseases, dermatologic diseases, allergic disorders, and respiratory diseases and for the palliative management of leukemia and lymphoma. Cortisone and hydrocortisone are less potent than prednisone and triamcinolone, but dexamethasone and betamethasone have the greatest anti-inflammatory potency. Disadvantages of corticosteroid use include the masking of signs of infection, an increase in the risk of peptic ulcer, and the development of edema and muscle weakness.

Manufacture of insulin

Insulin, secreted by the pancreas, is the principal hormone governing glucose metabolism. Insulin preparations were extracted from beef or pork pancreas until recombinant DNA technology made it possible to manufacture human insulin. Three preparations are available: rapid-acting (Regular, Semilente [trademark]), intermediate-acting (NPH, Lente [trademark]), and long-acting (PZI, Ultralente [trademark]). Other antidiabetic agents are available for treating non-insulin-dependent diabetes mellitus (NIDDM), also referred to as adult-onset diabetes, or type II diabetes. The sulfonylureas are oral hypoglycemic agents used as adjuncts to diet and exercise in the treatment of NIDDM.

Thyroid hormones include thyroxine and triiodothyronine, which regulate tissue metabolism. Natural desiccated thyroid produced from beef and pork and the synthetic derivatives levothyroxine and liothyronine are used in replacement therapy to treat hypothyroidism that results from any cause.

Drug therapy

GENERAL FEATURES

Principles of drug uptake and distribution. Study of the factors that influence the movement of drugs throughout the body is called pharmacokinetics, which includes the absorption, distribution, localization in tissues, biotransformation, and excretion of drugs. The study of the actions of the drugs and their effects is called pharmacodynamics. Before a drug can be effective, it must be absorbed and distributed throughout the body. Drugs taken orally may be absorbed by the intestines at different rates, some being absorbed rapidly, some more slowly. Even rapidly absorbed drugs can be prepared in ways that slow the degree of absorption and permit them to remain effective for 12 hours or longer. Drugs administered either intravenously or intramuscularly bypass problems of absorption, but dosage calculation is more critical.

Variation in drug response

Individuals respond differently to the same drug. Elderly persons, because of reduced kidney and liver function, may metabolize and excrete drugs more slowly. Because of this and other factors, the elderly usually require lower doses of medication than do younger people.

Other factors that affect the individual's response to drugs are the presence of disease, degree of nutrition or malnutrition, genetics, and the presence of other drugs in the system. Furthermore, just as the pain threshold varies among individuals, so does the response to drugs. Some people need higher-than-average doses; some, being very sensitive

to drugs, cannot tolerate even average doses, and they experience side effects when others do not.

Infants and children may have different rates of absorption than adults because bowel motility is irregular or gastric acidity is decreased. Drug distribution may be different in some people, such as premature infants who have little fatty tissue and a greater proportion of body water. Metabolic rates, which affect pharmacokinetics, are much higher during childhood, as anyone with a two-year-old can attest. The dosages of drugs for children are usually calculated on the basis of weight (milligrams per kilogram) or on the basis of body surface area (milligrams per square metre). If a drug has a wide margin of safety, it may be given as a fraction of the adult dose based on age, but the great variation in size among children of the same age complicates this computation. Children are not small adults, and drug dosages may be quite different than they are for adults.

The elderly are particularly susceptible to adverse drug effects because they often have multiple illnesses that require their taking various medications, some of which may be incompatible with others. In addition to decreased renal and hepatic function, gastric acid secretion decreases with age, and arteriosclerosis narrows the arteries, decreasing blood flow to the intestines and other organs. The precautions followed in prescribing medication for the elderly are an excellent example of the principle that should govern all drug therapy—drugs should be used in the lowest effective dose, especially because side effects increase with concentration. Because of illness or frailty, elderly people often have less reserve and may not be able to tolerate minor side effects that younger adults might not even notice.

When drugs are given in repeated doses, a steady state is achieved: the amount being administered equals the amount being excreted or metabolized. With some drugs, however, it may be difficult to determine the proper dose because of individual variations. In these cases, determining the plasma level of the drug may be useful, especially if the therapeutic window (*i.e.,* the concentration above which the drug is toxic and below which it is ineffective) is relatively small. Plasma levels of phenytoin, used to control epilepsy; digitalis, prescribed to combat heart failure; and lithium, used to moderate bipolar disorder, should be monitored.

Indications for use. The purpose of using drugs is to relieve symptoms, treat infection, reduce the risk of future disease, and destroy selected cells such as in the chemotherapeutic treatment of cancer. The best treatment, however, may not require a drug at all. Recognizing that no effective medication exists is just as important as knowing which one to select. When more than one drug is useful, physicians should select the one that is most effective, least hazardous, and least expensive. A recently developed drug may promise better results, yet it will be less predictable and possibly more expensive. Every drug has multiple actions: it will affect organs and systems beyond those to which it is specifically targeted. Some patients may also experience idiosyncratic effects (abnormal reactions peculiar to that individual) as well as allergic reactions to certain drugs—additional reasons to select drugs carefully and avoid their use altogether when simpler measures will work just as well. A case in point is the belief that penicillin or other antibiotics will cure viral infections—they will not. While new antiviral drugs are under development, using antibiotics unnecessarily is unwise and potentially dangerous. The number of drug-resistant organisms is growing and must be counteracted by the judicious prescribing of these chemicals.

Judicious prescription of medication

Unnecessary drug use also increases the possibility of drug interactions that may interfere with drug effectiveness. Interaction can occur in the stomach or intestinal tract where the presence of one drug may interfere with the absorption of another. Antacids, for example, reduce the absorption of the popular antibiotic tetracycline by forming insoluble complexes. Of greater importance is the interference of one drug with another. Some drugs can inhibit metabolism, which allows the amount of the drug to accumulate in the system, leading to potential toxicity if the dose is not decreased. Cimetidine, a drug used to treat peptic ulcers, causes few side effects by itself, but it does inhibit drug-metabolizing microsomal enzymes in the liver, increasing concentrations of many drugs that depend on these enzymes to be metabolized. This inhibition can be serious if the other drug is the anticoagulant warfarin. Bleeding can result if the dose is not reduced. Many other drugs are affected, such as antihypertensives (calcium channel blockers), antiarrhythmics (quinidine), and anticonvulsants (phenytoin). One drug can also decrease the renal excretion of another. Sometimes this effect is used to advantage, as, for example, when probenecid is given with penicillin to decrease its removal and thereby increase its concentration in the blood. But this type of interaction can be deadly: quinidine, for instance, can reduce the clearance of digoxin, a drug used to treat heart failure, potentially increasing its concentration to dangerous levels. Two drugs can also have additive effects, leading to toxicity, though either one alone would be therapeutic.

Problems with drug interactions can occur when a patient is being treated by different physicians, and one physician is not aware of the drug(s) that another has prescribed. Sometimes a physician may prescribe a drug to treat a symptom that actually is a side effect of another drug. Of course, discontinuing the first drug is preferable to adding another that may have side effects of its own. When a new symptom occurs, a recently initiated drug should be suspected before other causes are investigated. Patients should inform their physicians of any new drugs they are taking, as well as consult with the pharmacist about possible interactions that a nonprescription drug might have with a prescription drug already being taken. Having a personal physician who monitors all the drugs, both prescription and nonprescription, that the patient is taking is a wise course to follow.

Monitoring intake of drugs

In the United States, responsibility for assuring the safety and efficacy of prescription drugs is delegated to the Food and Drug Administration (FDA). This includes the approval of new drugs, identification of new indications, official labeling (to prevent unwarranted claims), surveillance of adverse drug reactions, and approval of methods of manufacture. Before an investigational new drug (IND) can be tested in humans, it must be submitted to and approved by the FDA. If clinical trials are successful, a new drug application (NDA) must be approved before it can be licensed and sold. This process usually takes years, but if the drug provides benefit to patients with life-threatening illnesses when existing treatments do not, then accelerated approval is possible. Physicians can receive permission to use an unapproved drug for a single patient. This consent, called emergency use and sometimes referred to as single-patient compassionate use, is granted if the situation is desperate and no other treatment is available. The FDA also sometimes grants approval to acquire drugs from other countries that are not available in the United States if a life-threatening situation seems to warrant this action. Another way to gain access to an investigational drug is to participate in a clinical trial. If it is a well-controlled, randomized, double-blind trial rather than an "open trial"—in which the investigator is not "blinded" and knows who is the subject and who is the control—the patient runs the risk of being given a placebo rather than the active drug.

The Federal Trade Commission (FTC) has responsibility for "truth in advertising" to assure that false or misleading claims are not made about foods, over-the-counter drugs, or cosmetics.

A rare disease presents a unique problem in treatment because the number of patients with the disease is so small (fewer than 200,000 in the United States) that it is not worthwhile for companies to go through the lengthy and expensive process required for approval and marketing. Drugs produced for such cases are made available under the Orphan Drug Act of 1983, which was intended to stimulate the development of drugs for rare diseases. More than 400 orphan drugs have been designated, but there are about 5,000 rare diseases that remain without treatment.

Controlled substances are drugs that foster dependence and have the potential for abuse. The Drug Enforcement Administration (DEA) regulates their manufacture, pre-

Controlled substances

scribing, and dispensing. Controlled substances are divided into five classes, or schedules, based on their potential for abuse or physical and psychological dependence. Schedule I encompasses heroin and other drugs with a high potential for abuse and no accepted medical use in the United States. Schedule II drugs, including narcotics such as opium and cocaine and stimulants such as amphetamines, have a high potential for abuse and dependence. Schedule III includes those drugs such as certain stimulants, depressants, barbiturates, and preparations containing limited amounts of codeine that cause moderate dependence. Schedule IV contains drugs that have limited potential for abuse or dependence, and includes some sedatives, antianxiety agents, and nonnarcotic analgesics. Schedule V drugs have an even lower potential for abuse than do schedule IV substances. Some, such as cough medicines and antidiarrheal agents containing limited amounts of codeine, can be purchased without a prescription. Physicians must have a DEA registration number to prescribe any controlled substance. Special triplicate prescription forms are required in certain states for schedule II drugs, and a patient's supply of these drugs cannot be replenished without a new prescription.

SYSTEMIC DRUG THERAPY

Systemic drug therapy involves treatment that affects the body as a whole or that acts specifically on systems that involve the entire body, such as the cardiovascular, respiratory, gastrointestinal, or nervous systems. Psychiatric disorders also are treated systemically.

Cholesterol deposits lead to atherosclerosis

The cardiovascular system. Atherosclerosis, the most common form of arteriosclerosis (generally called hardening of the arteries), is the thickening of large and medium-size arterial walls by cholesterol deposits that form plaques, causing the size of the arterial lumen to diminish. This narrowing compromises the artery's ability to supply blood to tissues and is most serious when the coronary arteries (those feeding the heart muscle) become clogged. A heart attack, with the death of a portion of the heart muscle, results; if the damage is extensive, sudden death will follow. The arteriosclerotic process can be slowed or even reversed by lowering serum cholesterol, especially the low-density lipoprotein (LDL) component. Cholesterol-reducing drugs, a low-cholesterol diet, exercise, and weight control can help. One form of cholesterol, high-density lipoprotein (HDL), is actually beneficial and helps to carry the harmful cholesterol out of the arterial wall. While some drugs will raise blood levels of high-density lipoprotein cholesterol, the most effective means of increasing it is to avoid cigarette smoke and increase exercise.

Narrowing of the coronary arteries can reduce the flow of blood to the heart and cause chest pain (angina pectoris). This condition can be treated with drugs such as nitroglycerin that primarily dilate the coronary arteries or by those such as the beta-blockers and calcium channel blockers that primarily reduce myocardial oxygen requirements.

Drugs that increase the strength of the heart muscle have been used to treat congestive heart failure for more than 200 years. Digitalis, derived from the foxglove plant, was the first drug found to have a positive inotropic effect (affects the force of muscular contraction) on the heart. Digoxin, the most commonly used form of this substance, can be given orally or intravenously. Digitalis has a relatively narrow therapeutic range: too much is toxic and can cause cardiac arrhythmias. Because toxicity is increased if the patient's serum potassium is low, close attention is paid to maintaining adequate potassium levels.

Drugs that dilate arterial smooth muscle and lower peripheral resistance (vasodilators) are also effective in treating heart failure by reducing the workload of the heart. The angiotensin converting enzyme (ACE) inhibitors are vasodilators used to treat heart failure. They also lower blood pressure in patients who are hypertensive.

The majority of cases of hypertension are due to unknown causes and are called essential, or primary, hypertension. Approximately 5 percent of all patients have secondary hypertension, which is high blood pressure that results from a known cause (*e.g.,* kidney disease). While the first treatment of hypertension should be to have the patient achieve normal weight, exercise, and reduce sodium in the diet, a wide variety of drugs are available to lower blood pressure, whether it be the systolic or diastolic measurement that is too high. A stepped-care approach has traditionally been used, starting with a single, well-tolerated drug, such as a diuretic. If it proves inadequate, a second drug is added and the combination manipulated until the most effective regimen with the fewest side effects is found. Occasionally, a third drug may be necessary.

Treatment of hypertension

The respiratory system. The drugs most frequently used for respiratory treatment are those that relieve cough in acute bronchitis. Antibiotics are effective only if the cause is bacterial. Most often, however, a virus is responsible, and the symptoms rather than the cause of the disease are treated, primarily with drugs that loosen or liquefy thick mucus (expectorants) and humidification (steam) that soothes the irritated mucous lining. While these treatments are widely prescribed, they have not been proven effective clinically. Cough suppressants are used to reduce unnecessary coughing but should not be employed excessively to subvert the cough's natural protective mechanism of ridding the airway of secretions and foreign substances. Dextromethorphan is a nonopioid cough suppressant nearly as effective as codeine and is available in over-the-counter preparations. If nasal congestion and postnasal drainage are present, an antihistamine and decongestant may be useful (see above *Treatment of symptoms: Cough*).

Asthma is a narrowing of the airways characterized by episodic wheezing. Bronchodilators are effective in a mild to moderate attack. Frequent attacks require long-term treatment with anti-inflammatory drugs such as cromolyn sodium, nedocromil sodium, or a corticosteroid.

Chronic obstructive pulmonary disease (COPD) manifests itself late in life with chronic cough and shortness of breath. Although most of the damage has already occurred, some benefit can still be obtained by stopping smoking, using bronchodilators, and administering antibiotics early when superimposed infection occurs. Supplemental oxygen therapy is used in severe cases.

The gastrointestinal system. Drugs are frequently used to reduce lower bowel activity when diarrhea occurs or to increase activity if constipation is the problem. Laxatives in the form of stimulants (cascara sagrada), bulk-forming agents (psyllium seed), osmotics (milk of magnesia), or lubricants (mineral oil) are commonly used. Diarrhea must be treated with appropriate antibiotics if the cause is bacterial, as in traveler's diarrhea, or with an antiparasitic agent if a parasite is to blame. Antidiarrheal agents include narcotics (codeine, paregoric), nonnarcotic analogs (loperamide hydrochloride), and bismuth subsalicylate (Pepto-Bismol [trademark]; see above *Treatment of symptoms: Diarrhea*).

Ulcers and chronic gastritis

Chronic gastritis and recurrent peptic ulcer often result from infection with *Helicobacter pylori* and are treated with antibiotics and bismuth. Ulcers not caused by *H. pylori* are treated with drugs that reduce the secretion of gastric acid, such as the H_2-receptor antagonists (cimetidine), or agents that form a barrier protecting the stomach against the acid (sucralfate). Antacids are used for additional symptomatic relief.

Nausea and vomiting are protective reflexes that should not be totally suppressed without the underlying cause being known. They may be psychogenic or caused by gastrointestinal or central nervous system disorders, medications, or systemic conditions (pregnancy or diabetic acidosis). Among the most widely used antiemetics are the phenothiazines (Compazine [trademark]), but new drugs continue to be developed that help control the vomiting related to cancer chemotherapy.

The nervous system. Alzheimer's disease is the most prevalent form of dementia (loss of intellectual function), and treatment had been primarily supportive until drugs that show modest promise for improving cognition (tacrine) were developed. Evidence that the continual use of cognitive faculties slows memory loss in the elderly has been supported by research showing that older persons who are stimulated regularly with memory exercises retain information better than those who are not.

Parkinsonism is named after James Parkinson, the English surgeon who in 1817 described "the shaking palsy." Although no treatment is known to halt the advance of the disease, levodopa and other drugs can significantly relieve the symptoms of tremor, muscular rigidity, and postural instability.

Migraine headache can be alleviated by one of the many forms of ergotamine and nonsteroidal anti-inflammatory drugs. Sumatriptan is a drug that has significantly improved the treatment of severe migraine attacks, causing fewer side effects than ergotamine or dihydroergotamine, but it is expensive.

Psychiatric disorders. Some of the greatest recent advances in pharmacotherapy have been in the treatment of anxiety disorders and depression. The benzodiazepines have been the mainstay of treatment for anxiety disorders since the 1960s, although their prolonged use incurs the risk of mild dependence. The azaspirodecanediones (buspirone) have little potential for producing dependency and are not affected by alcohol intake. Newer and safer medications are also available for treating panic disorder and obsessive-compulsive disorder.

Depression is among the most common life-threatening diseases, and considerable advances have been made in managing this very treatable disorder. The selective serotonin reuptake inhibitors (SSRIs) match previous antidepressants in effectiveness and have fewer unpleasant side effects. They also are safer if an overdose is taken, which is a significant threat in the case of severely depressed patients.

LOCAL DRUG THERAPY

Local anesthetics produce loss of sensation and make it possible for many surgical procedures to be performed without a general anesthetic. Barring any complications, the need for the patient to remain overnight in the hospital is obviated. Local anesthetics are also used to anesthetize specific peripheral nerves or larger nerve trunks. These nerve blocks can provide relief in painful conditions like rib fractures, but they are most frequently used to anesthetize an extremity during hand or foot surgery. Spinal anesthesia and epidural anesthesia, in which a local anesthetic is injected into the subarachnoid or epidural space of the lumbar (lower back) area of the spinal canal, provide pain relief during childbirth or surgery that involves the pelvic area yet lack the problems associated with a general anesthetic. Topical anesthetics, a type of local anesthetic, are also used on the skin, in the eye's conjunctiva and cornea, and in the mucous membranes of the nose, mouth, larynx, vagina, or urethra.

Topical anesthetics

Medications prescribed for dermatologic disorders account for a large amount of local drug therapy, whether it be a substance to stimulate hair growth or to soothe a burning and itching rash. Many different corticosteroid preparations are available to treat eczema, allergic reactions to substances like poison ivy, or seborrheic dermatitis. Sunblocks are used to protect the skin against ultraviolet rays and prevent skin cancer that can result from exposure to such radiation. Acne is controlled with skin cleansers, keratolytics to promote peeling, and topical antibiotics to prevent or treat infection. Physicians use various wet dressings, lotions, gels, creams, and ointments to treat acutely inflamed weeping and crusting sores and to moisturize and protect dry, cracked, and scaling skin. Burns heal more rapidly and with less scarring when treated appropriately with topical preparations like silver sulfadiazine. *Candida* infections of the mucous lining of the mouth (*i.e.,* thrush) or the vagina respond to nystatin or one of the imidazole drugs. The traditional treatment of genital warts has been the topical application of podophyllin, a crude resin, but new technology has made available interferon-α which is 70 percent effective when injected into the lesion itself or subcutaneously below it.

Most ophthalmic drugs are local—eye drops to treat glaucoma, steroid-antibacterial mixtures to treat infection, artificial tears for dry-eye syndromes, or mydriatics (drugs causing dilation of the pupil), like atropine, that facilitate refraction and internal examination of the eye.

CHEMOTHERAPY

Chemotherapy is the treatment of disease using chemical agents that are intended to eliminate the causative organism without harming the patient. In the strict sense, this applies to the use of antibiotics to treat such invading organisms as bacteria, viruses, fungi, or parasites. The term is commonly used, however, to describe the use of drugs to treat cancer, in which case the target is not a causative organism but wildly multiplying cells. The purpose of the therapy is to selectively kill tumour cells and leave normal cells unharmed—a very difficult task because most drugs have a narrow therapeutic zone beyond which they harm normal cells as well as cancer cells. Approximately 50 different anticancer drugs are available, and an equal number are currently being tested. Anticancer drugs are only relatively selective for cancer cells, and the toughest task for the physician is to select a drug that will destroy the most cancer cells, leave normal cells unharmed, and cause the fewest unpleasant and undesirable side effects. The therapeutic goal is to favourably balance the risk-benefit ratio in which the morbidity of the treatment is weighed against its potential benefits. If a treatment causes patients to be miserable and has only a slight chance of prolonging life, many patients will forego further treatment. However, if the potential for significantly prolonging survival by aggressive therapy exists, the patient may decide to continue with the therapy.

Cancer treatment

The effectiveness of chemotherapy depends on the highest possible concentration of the drug being at the tumour site sufficiently long to kill the tumour cells. The maximal opportunity for a cure exists in the early stage of the disease when the tumour is small and localized. The larger and more disseminated the tumour, the more difficult it is to eradicate. The stage the tumour is in will also determine the route of administration, which can be oral, intravenous, intra-abdominal, intrathecal (into the subarachnoid space of the spinal cord), or intra-arterial—specifically, into the artery feeding the tumour.

Suppression of bone marrow activity, which results in a decrease in blood cell production, represents the most limiting factor in chemotherapy. Because chemotherapy is most effective when used at the highest nontoxic dose, the interval between treatments may need to be prolonged to prevent complete bone marrow suppression. Supportive measures undertaken when bone marrow suppression occurs include repeated platelet transfusions (to combat bleeding caused by diminished platelet production) and white blood cell transfusions (to control infection).

Adjuvant chemotherapy is the use of drugs to eradicate or suppress residual disease after surgery or irradiation has been used to treat the tumour. This is necessary because distant micrometastases often occur beyond the primary tumour site. Adjuvant chemotherapy reduces the rate of recurrence of some cancers, especially ovarian cancer, osteogenic sarcoma, colon cancer, and Wilms' tumour. The antiestrogen drug tamoxifen has been effective in selected patients with breast cancer.

Surgical therapy

MAJOR CATEGORIES OF SURGERY

Wound treatment. Wounds, whether caused by accidental injury or a surgical scalpel, heal in three ways: (1) primary intention (wound edges are brought together, as in a clean surgical wound), (2) secondary intention (the wound is left open and heals by epithelization), or (3) third intention, or delayed closure (the wound is identified as potentially infected, is left open until contamination is minimized, and is then closed).

Choosing which method is best depends on whether excessive bacterial contamination is present, whether all necrotic material and foreign bodies can be identified and removed, and whether bleeding can be adequately controlled. Normal healing can occur only if the wound edges are clean and can be closely opposed without undue stress on the tissue. An adequate blood supply to the wound is essential. If the tissue is tight and the edges cannot be closed without tension, the blood supply will be compromised. Cutting under the skin to free it from the underly-

ing subcutaneous tissue may allow the edges to be brought together without tension. If direct approximation is still not possible, then skin grafts or flaps are used for closure.

Cleansing the wound

Wound closure begins with a thorough cleansing of the wound and the installation of a local anesthetic, usually lidocaine, which takes effect quickly and lasts for one to two hours. If the wound is contaminated, further cleansing is performed after instilling the local anesthetic, especially if foreign material is present. If the injury resulted from a fall on gravel or asphalt as in some motorcycle accidents, then aggressive scrubbing is needed to remove the many small pieces imbedded beneath the skin. High-pressure irrigation with saline solution will remove most foreign material and reduce the risk of subsequent infection. Contaminated wounds must be considered to be prone to infection with *Clostridium tetani,* which causes tetanus, and appropriate immunization should be given.

Sutures are the most commonly used means of wound closure, although staples and adhesive tissue tape may be more appropriate in certain circumstances. Silk sutures were originally used to close skin wounds, but nylon is stronger and causes less tissue reaction. Ideally, sutures are of the smallest possible diameter that will still maintain approximation of the wound edges. Absorbable sutures made of catgut (made not from cat but from sheep intestines) or a synthetic material such as polyglycolic acid are used to approximate the deeper layers of tissue beneath the skin so that tissue reaction will be lessened. The objective is to eliminate any unfilled space that could delay healing or allow fluid to accumulate. Drains connected to closed suction are used to prevent the collection of fluid when it is likely to accumulate, but drains serve as a source of contamination and are used infrequently. Staples permit faster closure of the skin but are less precise than sutures. When the edges can be brought together easily and without tension, tape is very useful. Although it is comfortable, easy to apply, and avoids the marks left by sutures, tape may come loose or be removed by the patient and is less successful if much wound edema occurs.

Sutures are removed after 3 to 14 days depending on the area involved, the cosmetic result desired, the blood supply to the area, and the amount of reaction that occurs around the sutures. Sutures on the face should be removed in 3 to 5 days to avoid suture marks. Tape is often used to provide support for the remainder of the time the wound needs to heal. Sutures on the trunk or leg will be removed after 7 to 10 days or longer if there is much tension on the wound. Tension and scarring are minimized in surgical procedures by making an incision parallel to normal skin lines, as in the horizontal neck incision for thyroidectomy.

Dressings protect the wound from external contamination and facilitate absorption of drainage. Because a surgical wound is most susceptible to surface contamination during the first 24 hours, an occlusive dressing is applied, consisting of gauze held in place by tape. Materials like transparent semipermeable membranes permit the wound to be observed without removal of the dressing and exposure of the wound to contamination. Dressings support the wound and, by adding compression, aid healing, as skin grafts do.

Wound healing

The healing of a wound results in scar formation; a strong yet minimally apparent scar is desirable. In some individuals a keloid, or thick overgrowth of scar, occurs no matter how carefully the wound was closed. The four phases of wound healing are inflammatory, migratory, proliferative, and late. The first, or inflammatory, phase occurs in the first 24 hours when platelets form a plug by adhering to the collagen exposed by damage to blood vessels. Fibrin joins the platelets to form a clot, and white blood cells invade the area to remove contamination by foreign material. Local blood vessels dilate to increase blood supply to the area, which hastens healing. In the second, or migratory, phase, fibroblasts and macrophages infiltrate the wound to initiate reconstruction. Capillaries grow in from the periphery, and epithelial cells advance across the clot to form a scab. In the proliferative phase, the fibroblasts produce collagen that increases wound strength, new epithelial cells cover the wound area, and capillaries join to form new blood vessels. In the late phase, the production of new and stronger collagen remodels the scar, blood vessels enlarge, and the epithelium at the surface heals.

Many factors, including diabetes mellitus or medications, can affect wound healing. In a patient whose diabetes is well controlled, wound healing is essentially normal, but, if the blood glucose level is elevated, it can impair healing and predispose the wound to infection. Kidney or liver failure and malnutrition also will delay wound healing, as will poor circulation owing to arteriosclerosis. Having steroids or anticancer or other drugs in the system can adversely affect the normal healing process.

Surgical extirpation. Extirpation is the complete removal or eradication of an organ or tissue and is a term usually used in cancer treatment or in the treatment of otherwise diseased or infected organs. The aim is to completely remove all cancerous tissue, which usually involves removing the visible tumour plus adjacent tissue that may contain microscopic extensions of the tumour. Excising a rim of adjacent, seemingly normal tissue ensures a complete cure unless there has been extension through the lymphatic system, which is the primary route for cancer to spread. For this reason, local lymph nodes are often removed with the tumour. Pathological examination of the nodes will show whether the cancer has spread. This indicates the likelihood of cure and whether additional treatment such as radiation or chemotherapy is needed. If complete removal of a tumour is not possible, palliative surgery, which provides relief but is not a cure, may be useful to relieve pain or pressure on adjacent structures. Radical surgery may not always be best, as in the early stages of breast cancer. Removal of the entire breast and surrounding structures, including the axillary lymph nodes, has been shown to provide no greater benefit than a lumpectomy (removal of the tumour only) followed by radiation to the area in early stages of breast cancer, while it often causes the patient increased psychological distress. However, because of improvements in breast reconstruction techniques, the trauma of a radical mastectomy is becoming less severe.

Palliative surgery

Reconstructive surgery. Reconstructive surgery is employed when a significant amount of tissue is missing as a result of trauma or surgical removal. A skin graft may be required if the wound cannot be closed directly. If a large surface area is involved, a thin split-thickness skin graft, consisting of epidermis only, is used. Unfortunately, although these grafts survive transplantation more successfully and heal more rapidly than other types of grafts, they are aesthetically displeasing because their appearance differs markedly from that of normal skin. In a small defect, especially one involving the face or hand, a full-thickness skin graft, consisting of epidermis and dermis, is used, and skin is generally donated from the ear, neck, or groin. Exposure of bone, nerve, or tendon requires a skin flap. This can be a local flap, in which tissue is freed and rotated from an adjacent area to cover the defect, or a free flap, in which tissue from another area of the body is used. An example of a local flap is the rotation of adjacent tissue (skin and subcutaneous tissue) to cover the defect left from removing a skin cancer. A free flap is used when the amount of tissue needed is not available locally, as in an injury to the lower leg from an automobile bumper. The amount and type of tissue needed and the blood supply available determine the type of flap to be used. The blood supply must be adequate to supply the separated flap and wound edge with nourishment.

Tissue expanders are another way of creating extra tissue that can be used to cover a defect. Inflatable plastic reservoirs are implanted under the normal skin of an adjacent area. For several weeks the reservoir is expanded with saline to stretch the overlying skin, which is then used to cover the defect.

Reconstructive surgery is performed for a variety of surgical conditions. It may require the fashioning of a new "organ," as in an artificial bladder, or may involve insertion of prosthetic devices such as artificial heart valves, pacemakers, joints, blood vessels, or bones.

Prosthetic devices can be used to replace diseased tissue. They usually perform better than donated tissue because

Prosthetic devices

they are made of material that does not stimulate rejection. The first prosthetic device to be used was the Dacron aortic graft developed by Michael E. De Bakey in 1954 to replace aortic aneurysms (dilated vessels that risk rupture and death) or vessels obstructed by arteriosclerotic plaques. Grafts made of similar materials are now employed to replace diseased arteries throughout the body. Other prosthetic devices include heart valves (made of plastic or taken from a pig) and metal joints (*e.g.,* hip, knee, or shoulder).

Transplantation surgery. The success of organ transplantation has greatly improved since the advent of the immunosuppressive drug cyclosporine. New and improved immunosuppressive drugs are currently being developed.

Kidney transplants are the most common, with those donated from living relatives ensuring the greatest prospects of long-term survival. The best survival rates are between identical twins. Cadaver transplants are often used, and one-year graft survival rate is 75 to 90 percent. Approximately 50 percent of grafts cease to function after 8 to 11 years, but others last 20 years or more. Kidneys removed from living donors can be preserved for up to 72 hours before they must be implanted, but most are implanted within 24 hours because successful transplantation decreases with time.

Heart and heart-lung organs can be preserved for four to six hours, and the success rate with this procedure continues to improve. Extensive blood and tissue matching is performed to minimize the risk of rejection. The size of the donor and donated organ should match the size of the recipient and the recipient's organ, and the time between pronouncement of death and procurement of the organ should be kept as short as possible.

In selected patients, liver transplantation has become an accepted treatment for end-stage liver disease. Mortality following surgery is 10 to 20 percent, and survivors still require long-term immunosuppressive therapy.

SURGICAL TECHNIQUES

Laser surgery. A laser is a device that produces an extremely intense monochromatic, nondivergent beam of light capable of generating intense heat when focused at close range. Its applications in the medical field include the surgical welding of a detached retina and the stanching of bleeding (called laser photocoagulation) in the gastrointestinal tract that can result from a peptic ulcer. Because a laser beam is absorbed by pigmented lesions, it can be used to treat pigmented tumours, remove tattoos, or coagulate a hemangioma (a benign but disfiguring tumour of blood vessels). Laser surgery has also been found to be effective in treating superficial bladder cancer and can be combined with ultrasonography for transurethral ultrasound-guided laser-induced prostatectomy (TULIP). More recent uses include the treatment of glaucoma and lesions of the cervix and vulva, including carcinoma in situ and genital warts.

Laser-induced prostatectomy

Cryosurgery. Cryosurgery is the destruction of tissue using extreme cold. Warts, precancerous skin lesions (actinic keratoses), and small cancerous skin lesions can be treated using liquid nitrogen. Other applications include removing cataracts, extirpating central nervous system lesions (including hard-to-reach brain tumours), and treating some heart conduction disorders.

Stereotaxic surgery. Stereotaxis (precise positioning in space) is a valuable neurosurgical technique that enables lesions deep in the brain that cannot be reached otherwise to be located and treated using cold (as in cryosurgery), heat, or chemicals. In this procedure, the head is held motionless in a head ring (halo frame), and the lesion or area to be treated is located using three-dimensional coordinates based on information from X rays and electrodes.

Stereotaxic techniques are also used to focus high-intensity radiation on localized areas of the brain to treat tumours or to obliterate arteriovenous malformations. This technique is also employed to guide fine-needle aspiration biopsies of brain lesions; it requires that only one burr hole be made in the skull with the patient under local anesthesia. Stereotaxic fine-needle biopsy also is used to evaluate breast lesions that are not palpable but are detected by mammography.

Minimally invasive surgery. Traditional open surgical techniques are being replaced by new technology in which a small incision is made and a rigid or flexible endoscope is inserted, enabling internal video imaging. Endoscopic procedures are commonly performed on nasal sinuses, intervertebral disks, fallopian tubes, shoulders, and knee joints, as well as on the gall bladder, appendix, and uterus. Although it has many advantages over traditional surgery, endosurgery may be more expensive and have higher complication rates than traditional approaches.

Trauma surgery. Trauma is one of the leading causes of loss of potential years of life. The explosion in the development of medical instrumentation and technology has made it possible for surgeons to save more lives than ever before thought possible. The intensive care unit contains a complex assortment of monitors and life-support equipment that can sustain life in situations that previously proved fatal, such as adult respiratory distress syndrome, multiorgan failure, kidney failure, and sepsis.

Radiation and other nonsurgical therapies

RADIATION THERAPY

Ionizing radiation is the transmission of energy by electromagnetic waves (*e.g.,* X rays) or by particles such as electrons, neutrons, or protons. Interaction with tissue produces free radicals and oxidants that damage or break cellular DNA, leading to cell death. When used properly, radiation may cause less damage than surgery and can often preserve organ structure and function. The type of radiation used depends on the radiosensitivity of the tumour and which healthy organs are within the radiation field. High-energy sources, such as linear accelerators, deposit their energy at a greater depth, sparing the skin but treating the deep-seated tumour. The radiation beam can also come from multiple directions, each beam being focused on the deep tumour, delivering a smaller dose to surrounding organs and tissues. Electron-beam radiation has low penetration and is useful in treating some skin cancers.

Radiation versus surgery

The basic unit of absorbed radiation is the gray (Gy): one gray equals 100 rads. Healthy organs have varying tolerance thresholds to radiation, bone marrow being the most sensitive and skin the least. The nervous system can tolerate much more radiation than the lungs or kidneys. Total body irradiation with approximately 10 Gy causes complete cessation of development of the bone marrow, and physicians use it to destroy defective tissue before performing a bone marrow transplant.

Radiation therapy can also be palliative if a cure is not possible; the size of the tumour can be reduced, thereby relieving pain or pressure on adjacent vital structures. It also can shrink a tumour to allow better drainage of an area, such as the lung, which can help to prevent infection and decrease the chance of bleeding.

Radioactive implants in the form of metal needles or "seeds" are used to treat some cancers, such as those of the prostate and uterine cervix. They can deliver high doses of radiation directly into the tumour with less effect on distant tissues.

An organ can also be irradiated by the ingestion of a radioactive substance. Hyperthyroidism can be treated with iodine-131, which collects in the thyroid gland and destroys a percentage of glandular tissue, thereby reducing function to normal. The drawback to this procedure is the difficulty in calculating the correct dose.

Irradiation is less effective in treating tissues that are poorly oxygenated (hypoxic) because of inadequate blood supply than it is in treating those that are well oxygenated. Some drugs enhance the toxic effect of radiation on tumour cells, especially those that are hypoxic.

OTHER NONINVASIVE THERAPIES

Hyperthermia. Some tumours are more sensitive than the surrounding healthy tissue to temperatures around 43° C (109.4° F). Sensitivity to heat is increased in the centre of tumours, where the blood supply is poor and radiation is less effective. A tumour may be heated using microwaves or ultrasound. Hyperthermia may enhance the effect of both radiation and chemotherapy; it is one form of nonionizing radiation therapy.

Nonionizing radiation therapy

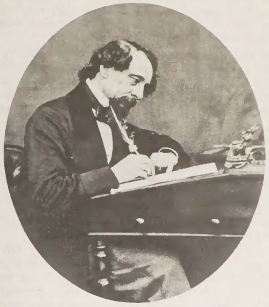
Photodynamic therapy. Another form of nonionizing radiation therapy is photodynamic therapy (PDT). This experimental technique involves administering a light-absorbing substance that is selectively retained by the tumour cells. The cells are killed by exposure to intense light, usually laser beams of appropriate wavelengths. Lesions amenable to PDT include tumours of the bronchus, bladder, skin, and peritoneal cavity.

Extracorporeal shock wave lithotripsy. The use of focused shock waves to pulverize stones in the urinary tract, usually the kidney or upper ureter, is called extracorporeal shock wave lithotripsy (ESWL). The resultant stone fragments or dust particles are passed through the ureter into the bladder and out the urethra. The patient is given a general, regional, or sometimes even local anesthetic and is immersed in water, and the shock wave is applied to the flank over the kidney. If the stone is small, submersion in a water bath is not necessary; shock waves are transmitted through the skin via a water-filled rubber bulb positioned over the stone site. Stones that are too large to be treated in this manner are removed by passing an endoscope into the ureter.

BIBLIOGRAPHY. RALPH H. MAJOR, *A History of Medicine,* 2 vol. (1954), covers medical history from its beginnings to modern times; unlike many history books, it is easy reading. MARK H. SWARTZ, *Textbook of Physical Diagnosis: History and Examination,* 2nd ed. (1994), an excellent illustrated text, covers the techniques of physical diagnosis. PAUL EKMAN and WALLACE V. FRIESEN, *Unmasking the Face: A Guide to Recognizing Emotions from Facial Clues* (1975, reprinted 1984), is a classic text in facial expression and emotion that uses composite photographs to show the importance of such areas as the brow, eyes, or mouth. AMERICAN PSYCHIATRIC ASSOCIATION, *Diagnostic and Statistical Manual of Mental Disorders: DSM-IV,* 4th ed. (1994), the standard reference, contains the diagnostic criteria for mental diseases as determined by the American Psychiatric Association. PAUL CUTLER, *Problem Solving in Clinical Medicine: From Data to Diagnosis,* 2nd ed. (1985), covers the fundamentals of problem solving and includes many examples. An unusual reference containing technical information not found in standard medical dictionaries is JAMES L. BENNINGTON, *Dictionary & Encyclopedia of Laboratory Medicine and Technology* (1984). ROBERT E. RAKEL, *Textbook of Family Practice,* 4th ed. (1990), is the standard textbook for family physicians covering the breadth of the discipline and emphasizing clinical diagnosis and treatment. A handy pocket reference presented in outline format containing diagnostic essentials for most medical conditions is DAVID C. DUGDALE and MICKEY S. EISENBERG, *Medical Diagnostics* (1992). ROBERT R. EDELMAN and STEVEN WARACH, "Magnetic Resonance Imaging," *The New England Journal of Medicine,* 328(10):708–716 (Mar. 11, 1993) and 328(11):785–791 (Mar. 18, 1993), provide a complete discussion of MRI including physical principles, uses, and cost-benefit considerations. Annually an issue of *JAMA,* the journal of the American Medical Association, is devoted to recent discoveries in every medical discipline and is an excellent source of up-to-date information on new developments in medicine; one such development is treated in RICHARD C. REBA, "Nuclear Medicine," *JAMA,* 270(2):230–232 (July 14, 1993). An expert panel reviewed the scientific literature and developed recommendations for prevention in 60 conditions that represent the leading causes of death and disability in the United States; these recommendations are found in U.S. PREVENTIVE SERVICES TASK FORCE, *Guide to Clinical Preventive Services* (1989). *Manual of Clinical Dietetics,* 4th ed. (1992), published by the American Dietetic Association, is an excellent reference for diets and nutritional contents of foods. *Drug Facts and Comparisons* (annual), contains detailed information about all prescription drugs. *Drug Evaluations Annual* is a well-written evaluation of the clinical use of specific drugs, including comparative evaluations. *Conn's Current Therapy* (annual), provides a concise reference for the treatment of most medical and surgical diseases. *Scientific American Medicine* (monthly), published in loose-leaf format, covers all major areas of internal medicine. DOUGLAS WILMORE (ed.), *Care of the Surgical Patient,* 2 vol. (1989), is a regularly updated loose-leaf publication from the Committee on Pre and Postoperative Care of the American College of Surgeons; vol. 1 is devoted to critical care and vol. 2 to elective surgery. *Current Medical Diagnosis & Treatment* (annual) contains concise diagnostic information and treatment for a large number of medical diseases. JAMES B. WYNGAARDEN, LLOYD J. SMITH, JR., and J. CLAUDE BENNETT, *Cecil Textbook of Medicine,* 19th ed. (1992), is one of the best standard textbooks of medicine, compiled by leading authorities and containing thorough information on common and rare diseases. RICHARD D. DESHAZO and DAVID L. SMITH (eds.), "Primer of Allergic and Immunologic Diseases," *JAMA,* 268(20):2785–2996 (Nov. 25, 1992), is an entire issue devoted to allergy and immunology, containing articles ranging from the basics of the immune response to autoimmune diseases and immunization; prepared by the American Academy of Allergy and Immunology, it provides complete coverage of the subject. (R.E.R.)

Dickens

Generally regarded as the greatest English novelist, Charles Dickens enjoyed a wider popularity than any previous author had done during his lifetime. Much in his work could appeal to simple and sophisticated, to the poor and to the Queen, and technological developments as well as the qualities of his work enabled his fame to spread worldwide very quickly. His long career saw fluctuations in the reception and sales of individual novels, but none of them was negligible or uncharacteristic or disregarded, and, though he is now admired for aspects and phases of his work that were given less weight by his contemporaries, his popularity has never ceased and his present critical standing is higher than ever before. The most abundantly comic of English authors, he was much more than a great entertainer. The range, compassion, and intelligence of his apprehension of his society and its shortcomings enriched his novels and made him both one of the great forces in 19th-century literature and an influential spokesman of the conscience of his age.



Dickens, 1859.
By courtesy of the Gernsheim Collection, the University of Texas at Austin

EARLY YEARS

Charles John Huffam Dickens was born February 7, 1812, in Portsmouth, Hampshire, but left it in infancy. His happiest childhood years were spent in Chatham (1817–22), an area to which he often reverts in his fiction. From 1822 he lived in London, until, in 1860, he moved permanently to a country house, Gad's Hill, near Chatham. His origins were middle class, if of a newfound and precarious respectability; one grandfather had been a domestic servant, and the other an embezzler. His father, a clerk in the navy pay office, was well paid, but his extravagance and ineptitude often brought the family to financial embarrassment or disaster. (Some of his failings and his ebullience are dramatized in Mr. Micawber in the partly autobiographical *David Copperfield*.) In 1824 the family reached bottom. Charles, the eldest son, had been withdrawn from school and was now set to manual work in a factory, and his father went to prison for debt. These shocks deeply affected Charles. Though abhorring this brief descent into the working class, he began to gain that sympathetic knowledge of their life and privations that informed his writings. Also, the images of the prison and of the lost, oppressed, or bewildered child recur in many novels. Much else in his character and art stems from this period, including, as the 20th-century novelist Angus Wilson has argued, his later difficulty, as man and author, in understanding women: this may be traced to his bitter resentment against his mother, who had, he felt, failed disastrously at this time to appreciate his sufferings. She had wanted him to stay at work when his father's release from prison and an improvement in the family's fortunes made the boy's return to school possible. Happily the father's view prevailed.

Factory experiences

His schooling, interrupted and unimpressive, ended at 15. He became a clerk in a solicitor's office, then a shorthand reporter in the lawcourts (thus gaining a knowledge of the legal world often used in the novels), and finally, like other members of his family, a parliamentary and newspaper reporter. These years left him with a lasting affection for journalism and contempt both for the law and for Parliament. His coming to manhood in the reformist 1830s, and particularly his working on the Liberal Benthamite *Morning Chronicle* (1834–36), greatly affected his political outlook. Another influential event now was his rejection as suitor to Maria Beadnell because his family and prospects were unsatisfactory; his hopes of gaining and chagrin at losing her sharpened his determination to succeed. His feelings about Maria then and at her later brief and disillusioning reentry into his life are reflected in David Copperfield's adoration of Dora Spenlow and in the middle-aged Arthur Clennam's discovery (in *Little Dorrit*) that Flora Finching, who had seemed enchanting years ago, was "diffuse and silly," that Flora "whom he had left a lily, had become a peony."

Early stories and essays

Beginning of literary career. Much drawn to the theatre, Dickens nearly became a professional actor in 1832. In 1833 he began contributing stories and descriptive essays to magazines and newspapers; these attracted attention and were reprinted as *Sketches by "Boz"* (February 1836). The same month, he was invited to provide a comic serial narrative to accompany engravings by a well-known artist; seven weeks later the first installment of *Pickwick Papers* appeared. Within a few months *Pickwick* was the rage and Dickens the most popular author of the day. During 1836 he also wrote two plays and a pamphlet on a topical issue (how the poor should be allowed to enjoy the Sabbath) and, resigning from his newspaper job, undertook to edit a monthly magazine, *Bentley's Miscellany*, in which he serialized *Oliver Twist* (1837–39). Thus, he had two serial installments to write every month. Already the first of his nine surviving children had been born; he had married (in April 1836) Catherine, eldest daughter of a respected Scottish journalist and man of letters, George Hogarth.

American tour

For several years his life continued at this intensity. Finding serialization congenial and profitable, he repeated the *Pickwick* pattern of 20 monthly parts in *Nicholas Nickleby* (1838–39); then he experimented with shorter weekly installments for *The Old Curiosity Shop* (1840–41) and *Barnaby Rudge* (1841). Exhausted at last, he then took a five-month vacation in America, touring strenuously and receiving quasi-royal honours as a literary celebrity but offending national sensibilities by protesting against the absence of copyright protection. A radical critic of British institutions, he had expected more from "the republic of my imagination," but he found more vulgarity and sharp practice to detest than social arrangements to admire. Some of these feelings appear in *American Notes* (1842) and *Martin Chuzzlewit* (1843–44).

First novels. His writing during these prolific years was remarkably various and, except for his plays, resourceful. *Pickwick* began as high-spirited farce and contained many conventional comic butts and traditional jokes; like other early works, it was manifestly indebted to the contemporary theatre, the 18th-century English novelists, and a few foreign classics, notably *Don Quixote*. But, besides giving new life to old stereotypes, *Pickwick* displayed, if sometimes in embryo, many of the features that were to

be blended in varying proportions throughout his fiction: attacks, satirical or denunciatory, on social evils and inadequate institutions; topical references; an encyclopaedic knowledge of London (always his predominant fictional locale); pathos; a vein of the macabre; a delight in the demotic joys of Christmas; a pervasive spirit of benevolence and geniality; inexhaustible powers of character creation; a wonderful ear for characteristic speech, often imaginatively heightened; a strong narrative impulse; and a prose style that, if here overdependent on a few comic mannerisms, was highly individual and inventive. Rapidly improvised and written only weeks or days ahead of its serial publication, *Pickwick* contains weak and jejune passages and is an unsatisfactory whole—partly because Dickens was rapidly developing his craft as a novelist while writing and publishing it. What is remarkable is that a first novel, written in such circumstances, not only established him overnight and created a new tradition of popular literature but also survived, despite its crudities, as one of the best known novels in the world.

Popularity of *Pickwick Papers*

His self-assurance and artistic ambitiousness had appeared in *Oliver Twist,* where he rejected the temptation to repeat the successful *Pickwick* formula. Though containing much comedy still, *Oliver Twist* is more centrally concerned with social and moral evil (the workhouse and the criminal world); it culminates in Bill Sikes's murdering Nancy and Fagin's last night in the condemned cell at Newgate. The latter episode was memorably depicted in George Cruikshank's engraving; the imaginative potency of Dickens' characters and settings owes much, indeed, to his original illustrators (Cruikshank for *Sketches by "Boz"* and *Oliver Twist,* "Phiz" [Hablot K. Browne] for most of the other novels until the 1860s). The currency of his fiction owed much, too, to its being so easy to adapt into effective stage versions. Sometimes 20 London theatres simultaneously were producing adaptations of his latest story; so even nonreaders became acquainted with simplified versions of his works. The theatre was often a subject of his fiction, too, as in the Crummles troupe in *Nicholas Nickleby.* This novel reverted to the *Pickwick* shape and atmosphere, though the indictment of the brutal Yorkshire schools (Dotheboys Hall) continued the important innovation in English fiction seen in *Oliver Twist*—the spectacle of the lost or oppressed child as an occasion for pathos and social criticism. This was amplified in *The Old Curiosity Shop,* where the death of Little Nell was found overwhelmingly powerful at the time, though a few decades later it became a byword for "Victorian sentimentality." In *Barnaby Rudge* he attempted another genre, the historical novel. Like his later attempt in this kind, *A Tale of Two Cities,* it was set in the late 18th century and presented with great vigour and understanding (and some ambivalence of attitude) the spectacle of large-scale mob violence.

To create an artistic unity out of the wide range of moods and materials included in every novel, with often several complicated plots involving scores of characters, was made even more difficult by Dickens' writing and publishing them serially. In *Martin Chuzzlewit* he tried "to resist the temptation of the current Monthly Number, and to keep a steadier eye upon the general purpose and design" (1844 Preface). Its American episodes had, however, been unpremeditated (he suddenly decided to boost the disappointing sales by some America-baiting and to revenge himself against insults and injuries from the American press). A concentration on "the general purpose and design" was more effective in the next novel, *Dombey and Son* (1846–48), though the experience of writing the shorter, and unserialized, Christmas books had helped him obtain greater coherence.

A Christmas Carol (1843), suddenly conceived and written in a few weeks, was the first of these Christmas books (a new literary genre thus created incidentally). Tossed off while he was amply engaged in writing *Chuzzlewit,* it was an extraordinary achievement—the one great Christmas myth of modern literature. His view of life was later to be described or dismissed as "Christmas philosophy," and he himself spoke of "*Carol* philosophy" as the basis of a projected work. His "philosophy," never very elaborated, involved more than wanting the Christmas spirit to prevail throughout the year, but his great attachment to Christmas (in his family life as well as his writings) is indeed significant and has contributed to his popularity. "Dickens dead?" exclaimed a London costermonger's girl in 1870. "Then will Father Christmas die too?"—a tribute both to his association with Christmas and to the mythological status of the man as well as of his work. The *Carol* immediately entered the general consciousness; Thackeray, in a review, called it "a national benefit, and to every man and woman who reads it a personal kindness." Further Christmas books, essays, and stories followed annually (except in 1847) through 1867. None equalled the *Carol* in potency, though some achieved great immediate popularity. Cumulatively they represent a celebration of Christmas attempted by no other great author.

How he struck his contemporaries in these early years appears in R.H. Horne's *New Spirit of the Age* (1844). Dickens occupied the first and longest chapter, as

> . . . manifestly the product of his age . . . a genuine emanation from its aggregate and entire spirit. . . . He mixes extensively in society, and continually. Few public meetings in a benevolent cause are without him. He speaks effectively. . . . His influence upon his age is extensive—pleasurable, instructive, healthy, reformatory. . . .
>
> Mr. Dickens is, in private, very much what might be expected from his works. . . . His conversation is genial . . . [He] has singular personal activity, and is fond of games of practical skill. He is also a great walker, and very much given to dancing Sir Roger de Coverley. In private, the general impression of him is that of a first-rate practical intellect, with "no nonsense" about him.

He was indeed very much a public figure, actively and centrally involved in his world, and a man of confident presence. He was reckoned the best after-dinner speaker of the age; other superlatives he attracted included his having been the best shorthand reporter on the London press and his being the best amateur actor on the stage. Later he became one of the most successful periodical editors and the finest dramatic recitalist of the day. He was splendidly endowed with many skills. "Even irrespective of his literary genius," wrote an obituarist, "he was an able and strong-minded man, who would have succeeded in almost any profession to which he devoted himself" (*Times,* June 10, 1870). Few of his extraliterary skills and interests were irrelevant to the range and mode of his fiction.

Status as a public figure

Privately in these early years, he was both domestic and social. He loved home and family life and was a proud and efficient householder; he once contemplated writing a cookbook. To his many children, he was a devoted and delightful father, at least while they were young; relations with them proved less happy during their adolescence. Apart from periods in Italy (1844–45) and Switzerland and France (1846–47), he still lived in London, moving from an apartment in Furnival's Inn to larger houses as his income and family grew. Here he entertained his many friends, most of them popular authors, journalists, actors, or artists, though some came from the law and other professions or from commerce and a few from the aristocracy. Some friendships dating from his youth endured to the end, and, though often exasperated by the financial demands of his parents and other relatives, he was very fond of some of his family and loyal to most of the rest. Some literary squabbles came later, but he was on friendly terms with most of his fellow authors, of the older generation as well as his own. Necessarily solitary while writing and during the long walks (especially through the streets at night) that became essential to his creative processes, he was generally social at other times. He enjoyed society that was unpretentious and conversation that was genial and sensible but not too intellectualized or exclusively literary. High society he generally avoided, after a few early incursions into the great houses; he hated to be lionized or patronized.

He had about him "a sort of swell and overflow as of a prodigality of life," an American journalist said. Everyone was struck by the brilliance of his eyes and his smart, even dandyish, appearance ("I have the fondness of a savage for finery," he confessed). John Forster, his intimate friend and future biographer, recalled him at the *Pickwick* period:

the quickness, keenness, and practical power, the eager, restless, energetic outlook on each several feature [of his face] seemed to tell so little of a student or writer of books, and so much of a man of action and business in the world. Light and motion flashed from every part of it.

He was proud of his art and devoted to improving it and using it to good ends (his works would show, he wrote, that "Cheap Literature is not behind-hand with the Age, but holds its place, and strives to do its duty"), but his art never engaged all his formidable energies. He had no desire to be narrowly literary.

Unsuccessful journalistic ventures

A notable, though unsuccessful, demonstration of this was his being founder-editor in 1846 of the *Daily News* (soon to become the leading Liberal newspaper). His journalistic origins, his political convictions and readiness to act as a leader of opinion, and his wish to secure a steady income independent of his literary creativity and of any shifts in novel readers' tastes made him attempt or plan several periodical ventures in the 1840s. The return to daily journalism soon proved a mistake—the biggest fiasco in a career that included few such misdirections or failures. A more limited but happier exercise of his practical talents began soon afterward: for more than a decade he directed, energetically and with great insight and compassion, a reformatory home for young female delinquents, financed by his wealthy friend Angela Burdett-Coutts. The benevolent spirit apparent in his writings often found practical expression in his public speeches, fund-raising activities, and private acts of charity.

Dombey and Son (1846–48) was a crucial novel in his development, a product of more thorough planning and maturer thought and the first in which "a pervasive uneasiness about contemporary society takes the place of an intermittent concern with specific social wrongs" (Kathleen Tillotson). Using railways prominently and effectively, it was very up-to-date, though the questions posed included such perennial moral and religious challenges as are suggested by the child Paul's first words in the story: "Papa, what's money?" Some of the corruptions of money and pride of place and the limitations of "respectable" values are explored, virtue and human decency being discovered most often (as elsewhere in Dickens) among the poor, humble, and simple. In Paul's early death Dickens offered another famous pathetic episode; in Mr. Dombey he made a more ambitious attempt than before at serious and internal characterization. *David Copperfield* (1849–50) has been described as a "holiday" from these larger social concerns and most notable for its childhood chapters, "an enchanting vein which he had never quite found before and which he was never to find again" (Edmund Wilson). Largely for this reason and for its autobiographical interest, it has always been among his most popular novels and was Dickens' own "favourite child." It incorporates material from the autobiography he had recently begun but soon abandoned and is written in the first person, a new technique for him. David differs from his creator in many ways, however, though Dickens uses many early experiences that had meant much to him—his period of work in the factory while his father was jailed, his schooling and reading, his passion for Maria Beadnell, and (more cursorily) his emergence from parliamentary reporting into successful novel writing. In Micawber the novel presents one of the "Dickens characters" whose imaginative potency extends far beyond the narratives in which they figure; Pickwick and Sam Weller, Mrs. Gamp and Mr. Pecksniff, and Scrooge are some others.

Autobiographical interest of *David Copperfield*

MIDDLE YEARS

Journalism. Dickens' journalistic ambitions at last found a permanent form in *Household Words* (1850–59) and its successor, *All the Year Round* (1859–88). Popular weekly miscellanies of fiction, poetry, and essays on a wide range of topics, these had substantial and increasing circulations, reaching 300,000 for some of the Christmas numbers. Dickens contributed some serials—the lamentable *Child's History of England* (1851–53), *Hard Times* (1854), *A Tale of Two Cities* (1859), and *Great Expectations* (1860–61)—and essays, some of which were collected in *Reprinted Pieces* (1858) and *The Uncommercial Traveller* (1861, later amplified). Particularly in 1850–52 and during the Crimean War, he contributed many items on current political and social affairs; in later years he wrote less—much less on politics—and the magazine was less political, too. Other distinguished novelists contributed serials, including Mrs. Gaskell, Wilkie Collins, Charles Reade, and Bulwer Lytton. The poetry was uniformly feeble; Dickens was imperceptive here. The reportage, often solidly based, was bright (sometimes painfully so) in manner. His conduct of these weeklies shows his many skills as editor and journalist but also some limitations in his tastes and intellectual ambitions. The contents are revealing in relation to his novels: he took responsibility for all the opinions expressed (for articles were anonymous) and selected and amended contributions accordingly; thus comments on topical events and so on may generally be taken as representing his opinions, whether or not he wrote them. No English author of comparable status has devoted 20 years of his maturity to such unremitting editorial work, and the weeklies' success was due not only to his illustrious name but also to his practical sagacity and sustained industry. Even in his creative work, as his eldest son said,

> No city clerk was ever more methodical or orderly than he; no humdrum, monotonous, conventional task could ever have been discharged with more punctuality, or with more businesslike regularity.

Novels. The novels of these years, *Bleak House* (1852–53), *Hard Times* (1854), and *Little Dorrit* (1855–57), were much "darker" than their predecessors. Presenting a remarkably inclusive and increasingly sombre picture of contemporary society, they were inevitably often seen at the time as fictionalized propaganda about ephemeral issues. They are much more than this, though it is never easy to state how Dickens' imagination transforms their many topicalities into an artistically coherent vision that transcends their immediate historical context. Similar questions are raised by his often basing fictional characters, places, and institutions on actual originals. He once spoke of his mind's taking "a fanciful photograph" of a scene, and there is a continual interplay between photographic realism and "fancy" (or imagination). "He describes London like a special correspondent for posterity" (Walter Bagehot, 1858), and posterity has certainly found in his fiction the response of an acute, knowledgeable, and concerned observer to the social and political developments of "the moving age." In the novels of the 1850s, he is politically more despondent, emotionally more tragic. The satire is harsher, the humour less genial and abundant, the "happy endings" more subdued than in the early fiction. Technically, the later novels are more coherent, plots being more fully related to themes, and themes being often expressed through a more insistent use of imagery and symbols (grim symbols, too, such as the fog in *Bleak House* or the prison in *Little Dorrit*). His art here is more akin to poetry than to what is suggested by the photographic or journalistic comparisons. "Dickensian" characterization continues in the sharply defined and simplified grotesque or comic figures, such as Chadband in *Bleak House* or Mrs. Sparsit in *Hard Times*, but large-scale figures of this type are less frequent (the Gamps and Micawbers belong to the first half of his career). Characterization also has become more subordinate to "the general purpose and design"; moreover, Dickens is presenting characters of greater complexity, who provoke more complex responses in the reader (William Dorrit, for instance). Even the juvenile leads, who had usually been thinly conceived conventional figures, are now often more complicated in their make-up and less easily rewarded by good fortune. With his secular hopes diminishing, Dickens becomes more concerned with "the great final secret of all life"—a phrase from *Little Dorrit*, where the spiritual dimension of his work is most overt. Critics disagree as to how far so worldly a novelist succeeds artistically in enlarging his view to include the religious. These novels, too, being manifestly an ambitious attempt to explore the prospects of humanity at this time, raise questions, still much debated, about the intelligence and profundity of his understanding of society.

Themes of the later novels

Personal unhappiness. Dickens' spirits and confidence

in the future had indeed declined: 1855 was "a year of much unsettled discontent for him," his friend Forster recalled, partly for political reasons (or, as Forster hints, his political indignation was exacerbated by a "discontent" that had personal origins). The Crimean War, besides exposing governmental inefficiency, was distracting attention from the "poverty, hunger, and ignorant desperation" at home. In *Little Dorrit,* "I have been blowing off a little of indignant steam which would otherwise blow me up . . . ," he wrote, "but I have no present political faith or hope—not a grain." Not only were the present government and Parliament contemptible but "representative government is become altogether a failure with us, . . . the whole thing has broken down . . . and has no hope in it." Nor had he a coherent alternative to suggest. This desperation coincided with an acute state of personal unhappiness. The brief tragicomedy of Maria Beadnell's reentry into his life, in 1855, finally destroyed one nostalgic illusion and also betrayed a perilous emotional immaturity and hunger. He now openly identified himself with some of the sorrows dramatized in the adult David Copperfield:

> Why is it, that as with poor David, a sense comes always crushing on me, now, when I fall into low spirits, as of one happiness I have missed in life, and one friend and companion I have never made?

Marital grief

This comes from the correspondence with Forster in 1854–55, which contains the first admissions of his marital unhappiness; by 1856 he is writing, "I find the skeleton in my domestic closet is becoming a pretty big one"; by 1857–58, as Forster remarks, an "unsettled feeling" had become almost habitual with him, "and the satisfactions which home should have supplied, and which indeed were essential requirements of his nature, he had failed to find in his home." From May 1858, Catherine Dickens lived apart from him. A painful scandal arose, and Dickens did not act at this time with tact, patience, or consideration. The affair disrupted some of his friendships and narrowed his social circle, but surprisingly it seems not to have damaged his popularity with the public.

Catherine Dickens maintained a dignified silence, and most of Dickens' family and friends, including his official biographer, Forster, were discreetly reticent about the separation. Not until 1939 did one of his children (Katey), speaking posthumously through conversations recorded by a friend, offer a candid inside account. It was discreditable to him, and his self-justifying letters must be viewed with caution. He there dated the unhappiness of his marriage back to 1838, attributed to his wife various "peculiarities" of temperament (including her sometimes labouring under "a mental disorder"), emphatically agreed with her (alleged) statement that "she felt herself unfit for the life she had to lead as my wife," and maintained that she never cared for the children nor they for her. In more temperate letters, where he acknowledged her "amiable and complying" qualities, he simply and more acceptably asserted that their temperaments were utterly incompatible. She was, apparently, pleasant but rather limited; such faults as she had were rather negative than positive, though family tradition from a household that knew the Dickenses well speaks of her as "a whiney woman" and as having little understanding of, or patience with, the artistic temperament.

Friendship with Ellen Ternan

Dickens' self-justifying letters lack candour in omitting to mention Ellen Ternan, an actress 27 years his junior, his passion for whom had precipitated the separation. Two months earlier he had written more frankly to an intimate friend:

> The domestic unhappiness remains so strong upon me that I can't write, and (waking) can't rest, one minute. I have never known a moment's peace or content, since the last night of The Frozen Deep.

The Frozen Deep was a play in which he and Nelly (as Ellen was called) had performed together in August 1857. She was an intelligent girl, of an old theatrical family; reports speak of her as having "a pretty face and well-developed figure"—or "passably pretty and not much of an actress." She left the stage in 1860; after Dickens' death she married a clergyman and helped him run a school. The affair was hushed up until the 1930s, and evidence about it remains scanty, but every addition confirms that Dickens was deeply attached to her and that their relationship lasted until his death. It seems likely that she became his mistress, though probably not until the 1860s; assertions that a child, or children, resulted remain unproved. Similarly, suggestions that the anguish experienced by some of the lovers in the later novels may reflect Dickens' own feelings remain speculative. It is tempting, indeed, to associate Nelly with some of their heroines (who are more spirited and complex, less of the "legless angel," than most of their predecessors), especially as her given names, Ellen Lawless, seem to be echoed by those of heroines in the three final novels—Estella, Bella, and Helena Landless—but nothing definite is known about how she responded to Dickens, what she felt for him at the time, or how close any of these later love stories were to aspects or phases of their relationship.

"There is nothing very remarkable in the story," commented one early transmitter of it, and this seems just. Many middle-aged men feel an itch to renew their emotional lives with a pretty young girl, even if, unlike Dickens, they cannot plead indulgence for "the wayward and unsettled feeling which is part (I suppose) of the tenure on which one holds an imaginative life." But the eventual disclosure of this episode caused surprise, shock, or piquant satisfaction, being related of a man whose rebelliousness against his society had seemed to take only impeccably reformist shapes. A critic in 1851, listing the reasons for his unique popularity, had cited "above all, his deep reverence for the household sanctities, his enthusiastic worship of the household gods." After these disclosures he was, disconcertingly or intriguingly, a more complex man; and, partly as a consequence, Dickens the novelist also began to be seen as more complex, less conventional, than had been realized. The stimulus was important, though Nelly's significance, biographically and critically, has proved far from inexhaustible.

Public readings. In the longer term, Kathleen Tillotson's remark is more suggestive: "his lifelong love-affair with his reading public, when all is said, is by far the most interesting love-affair of his life." This took a new form, about the time of Dickens' separation from his wife, in his giving public readings from his works, and it is significant that, when trying to justify this enterprise as certain to succeed, he referred to "that particular relation (personally affectionate and like no other man's) which subsists between me and the public." The remark suggests how much Dickens valued his public's affection, not only as a stimulus to his creativity and a condition for his commercial success but also as a substitute for the love he could not find at home. He had been toying with the idea of turning paid reader since 1853, when he began giving occasional readings in aid of charity. The paid series began in April 1858, the immediate impulse being to find some energetic distraction from his marital unhappiness. But the readings drew on more permanent elements in him and his art: his remarkable histrionic talents, his love of theatricals and of seeing and delighting an audience, and the eminently performable nature of his fiction. Moreover, he could earn more by reading than by writing, and more certainly; it was easier to force himself to repeat a performance than create a book.

Histrionic talents

His initial repertoire consisted entirely of Christmas books but was soon amplified by episodes from the novels and magazine Christmas stories. A performance usually consisted of two items; of the 16 eventually performed, the most popular were "The Trial from *Pickwick*" and the *Carol.* Comedy predominated, though pathos was important in the repertoire, and horrifics were startlingly introduced in the last reading he devised, "Sikes and Nancy," with which he petrified his audiences and half killed himself. Intermittently, until shortly before his death, he gave seasons of readings in London and embarked upon hard-working tours through the provinces and (in 1867–68) the United States. Altogether he performed about 471 times. He was a magnificent performer, and important elements in his art—the oral and dramatic qualities—were demonstrated in these renderings. His insight and skill revealed nuances in the narration and characterization that few

readers had noticed. Necessarily, such extracts or short stories, suitable for a two-hour entertainment, excluded some of his larger and deeper effects—notably, his social criticism and analysis—and his later novels were underrepresented. Dickens never mentions these inadequacies. He manifestly enjoyed the experience until, near the end, he was becoming ill and exhausted. He was writing much less in the 1860s. It is debatable how far this was because the readings exhausted his energies, while providing the income, creative satisfaction, and continuous contact with an audience that he had formerly obtained through the novels. He gloried in his audiences' admiration and love. Some friends thought this too crude a gratification, too easy a triumph, and a sad declension into a lesser and ephemeral art. In whatever way the episode is judged, it was characteristic of him—of his relationship with his public, his business sense, his stamina, his ostentatious display of supplementary skills, and also of his originality. No important author (at least, according to reviewers, since Homer) and no English author since who has had anything like his stature has devoted so much time and energy to this activity. The only comparable figure is his contemporary, Mark Twain, who acknowledged Dickens as the pioneer.

LAST YEARS

Final novels. Tired and ailing though he was, he remained inventive and adventurous in his final novels. *A Tale of Two Cities* (1859) was an experiment, relying less than before on characterization, dialogue, and humour. An exciting and compact narrative, it lacks too many of his strengths to count among his major works. Sydney Carton's self-sacrifice was found deeply moving by Dickens and by many readers; Dr. Manette now seems a more impressive achievement in serious characterization. The French Revolution scenes are vivid, if superficial in historical understanding. *Great Expectations* (1860–61) resembles *Copperfield* in being a first-person narration and in drawing on parts of Dickens' personality and experience. Compact like its predecessor, it lacks the panoramic inclusiveness of *Bleak House, Little Dorrit,* and *Our Mutual Friend,* but, though not his most ambitious, it is his most finely achieved novel. The hero Pip's mind is explored with great subtlety, and his development through a childhood and youth beset with hard tests of character is traced critically but sympathetically. Various "great expectations" in the book proved ill founded—a comment as much on the values of the age as on the characters' weaknesses and misfortunes. *Our Mutual Friend* (1864–65), a large inclusive novel, continues this critique of monetary and class values. London is now grimmer than ever before, and the corruption, complacency, and superficiality of "respectable" society are fiercely attacked. Many new elements are introduced into Dickens' fictional world, but his handling of the old comic-eccentrics (such as Boffin, Wegg, and Venus) is sometimes tiresomely mechanical. How the unfinished *Edwin Drood* (1870) would have developed is uncertain. Here again Dickens left panoramic fiction to concentrate on a limited private action. The central figure was evidently to be John Jasper, whose eminent respectability as a cathedral organist was in extreme contrast to his haunting low opium dens and, out of violent sexual jealousy, murdering his nephew. It would have been his most elaborate treatment of the themes of crime, evil, and psychological abnormality that had recurred throughout his novels; a great celebrator of life, he was also obsessed with death.

Change in Dickens' personality

How greatly Dickens personally had changed appears in remarks by friends who met him again, after many years, during the American reading tour in 1867–68. "I sometimes think . . .," wrote one, "I must have known two individuals bearing the same name, at various periods of my own life." But just as the fiction, despite many developments, still contained many stylistic and narrative features continuous with the earlier work, so, too, the man remained a "human hurricane," though he had aged considerably, his health had deteriorated, and his nerves had been jangled by travelling ever since his being in a railway accident in 1865. Other Americans noted that, though grizzled, he was "as quick and elastic in his movements as ever." His photographs, wrote a journalist after one of the readings, "give no idea of his genial expression. To us he appears like a hearty, companionable man, with a deal of fun in him." But that very day Dickens was writing, "I am nearly used up," and listing the afflictions now "telling heavily upon me." His pride and the old-trouper tradition made him conceal his sufferings. And, if sometimes by an effort of will, his old high spirits were often on display. "The cheerfullest man of his age," he was called by his American publisher, J.T. Fields; Fields's wife more perceptively noted, "Wonderful, the flow of spirits C.D. has for a sad man."

His fame remained undiminished, though critical opinion was increasingly hostile to him. Henry Wadsworth Longfellow, noting the immense enthusiasm for him during the American tour, remarked: "One can hardly take in the whole truth about it, and feel the universality of his fame." But in many respects he was "a sad man" in these later years. He never was tranquil or relaxed. Various old friends were now estranged or dead or for other reasons less available; he was now leading a less social life and spending more time with young friends of a calibre inferior to his former circle. His sons were causing much worry and disappointment; "all his fame goes for nothing," said a friend, "since he has not the one thing. He is very unhappy in his children." His life was not all dreary, however. He loved his country house, Gad's Hill, and he could still "warm the social atmosphere wherever he appeared with that summer glow which seemed to attend him." T.A. Trollope (contributor to Dickens' *All the Year Round* and brother of the novelist Anthony Trollope), who wrote that, despaired of giving people who had not met him any idea of

> the general charm of his manner. . . . His laugh was brimful of enjoyment. . . . His enthusiasm was boundless. . . . He was a *hearty* man, a large-hearted man, . . . a strikingly manly man.

Final months

Farewell readings. His health remained precarious after the punishing American tour and was further impaired by his addiction to giving the strenuous "Sikes and Nancy" reading. His farewell reading tour was abandoned when, in April 1869, he collapsed. He began writing another novel and gave a short farewell season of readings in London, ending with the famous speech, "From these garish lights I vanish now for evermore . . ."—words repeated, less than three months later, on his funeral card. He died suddenly at Gad's Hill on June 9, 1870, and was buried in Westminster Abbey.

ASSESSMENT

Contemporary opinion. Ralph Waldo Emerson, attending one of Dickens' readings in Boston, "laughed as if he must crumble to pieces," but, discussing Dickens afterward, he said:

> I am afraid he has too much talent for his genius; it is a fearful locomotive to which he is bound and can never be free from it nor set to rest. . . . He daunts me! I have not the key.

There is no simple key to so prolific and multifarious an artist nor to the complexities of the man, and interpretation of both is made harder by his possessing and feeling the need to exercise so many talents besides his imagination. How his fiction is related to these talents—practical, journalistic, oratorical, histrionic—remains controversial. Also the geniality and unequalled comedy of the novels must be related to the sufferings, errors, and self-pity of their author and to his concern both for social evils and for the perennial griefs and limitations of humanity. The novels cover a wide range, social, moral, emotional, and psychological. Thus, he is much concerned with very ordinary people but also with abnormality (*e.g.*, eccentricity, depravity, madness, hallucinations, dream states). He is both the most imaginative and fantastic and the most topical and documentary of great novelists. He is unequal, too; a wonderfully inventive and poetic writer, he can also, even in his mature novels, write with a painfully slack conventionality.

Biographers have only since the mid-20th century known enough to explore the complexity of Dickens' nature. Critics have always been challenged by his art, though

Most popular novels

from the start it contained enough easily acceptable ingredients, evident skill and gusto, to ensure popularity. The earlier novels were and by and large have continued to be Dickens' most popular works: *The Pickwick Papers, Oliver Twist, Martin Chuzzlewit, A Christmas Carol,* and *David Copperfield.* Critics began to demur against the later novels, deploring the loss of the freer comic spirit, baffled by the more symbolic mode of his art, and uneasy when the simpler reformism over isolated issues became a more radical questioning of social assumptions and institutions. Dickens was never neglected or forgotten and never lost his popularity, but for 70 years after his death he received remarkably little serious attention (George Gissing, G.K. Chesterton, and George Bernard Shaw being notable exceptions). F.R. Leavis, later to revise his opinion, was speaking for many, in 1948, when he asserted that "the adult mind doesn't as a rule find in Dickens a challenge to an unusual and sustained seriousness"; Dickens was indeed a great genius, "but the genius was that of a great entertainer."

Modern criticism. Modern Dickens criticism dates from 1940–41, with the very different impulses given by George Orwell, Edmund Wilson, and Humphry House. In the 1950s, a substantial reassessment and re-editing of the works began, his finest artistry and greatest depth now being discovered in the later novels—*Bleak House, Little Dorrit,* and *Great Expectations*—and (less unanimously) in *Hard Times* and *Our Mutual Friend.* Scholars have explored his working methods, his relations with his public, and the ways in which he was simultaneously an eminently Victorian figure and an author "not of an age but for all time." Biographically, little had been added to Forster's massive and intelligent *Life* (1872–74), except the Ellen Ternan story, until Edgar Johnson's in 1952. Since then, no radically new view has emerged, though several works—including those by Joseph Gold (1972) and Fred Kaplan (1975)—have given particular phases or aspects fuller attention. The centenary in 1970 demonstrated a critical consensus about his standing second only to William Shakespeare in English literature, which would have seemed incredible 40 or even 20 years earlier.

MAJOR WORKS

NOVELS: *The Pickwick Papers* (1837); *Oliver Twist* (1838); *Nicholas Nickleby* (1839); *The Old Curiosity Shop* and *Barnaby Rudge* (1841), two novels first published in a "clock framework," later abandoned, under the title of *Master Humphrey's Clock; Martin Chuzzlewit* (1844); *Dombey and Son* (1848); *David Copperfield* (1850); *Bleak House* (1853); *Hard Times* (1854); *Little Dorrit* (1857); *A Tale of Two Cities* (1859); *Great Expectations* (1861); *Our Mutual Friend* (1865); *The Mystery of Edwin Drood* (1870, unfinished).

CHRISTMAS BOOKS: *A Christmas Carol* (1843); *The Chimes* (1845, for 1844); *The Cricket on the Hearth* (1846, for 1845); *The Battle of Life* (1846); *The Haunted Man* (1848).

STORIES (CHRISTMAS STORIES): The volume entitled *Christmas Stories* in collected editions includes "A Christmas Tree" (1850); "What Christmas Is as We Grow Older" (1851); "The Poor Relation's Story" (1852); "Nobody's Story" (1853); "The Seven Poor Travellers" (1854); "The Holly-Tree," sometimes called "The Holly-Tree Inn" (1855); "The Wreck of the Golden Mary" (1856); "The Perils of Certain English Prisoners" (1857); "Going into Society" (1858); "The Haunted House" (1859); "A Message from the Sea" (1860); "Tom Tiddler's Ground" (1861); "Somebody's Luggage" (1862); "Mrs. Lirriper's Lodgings" (1863); "Mrs. Lirriper's Legacy" (1864); "Doctor Marigold" (1865); "Mugby Junction" (1866); "No Thoroughfare" (1867).

OTHER STORIES): in collected editions generally appended to the volume entitled *Reprinted Pieces,* ["The Lamplighter" (1841);] "To Be Read at Dusk" (1852); "Hunted Down" (1859); "George Silverman's Explanation" (1867); "Holiday Romance" (1868; children's story in 4 parts; pt. 2, "The Magic Fishbone," often reprinted separately).

OTHER WORKS: *Sketches by "Boz,"* 2 series (1836, together, 1839, included Dickens' first published work, "A Dinner at Poplar Walk," 1833); *Sketches of Young Gentlemen* (1838) and *Sketches of Young Couples* (1840), both usually appended to the *Sketches by "Boz"* volume, in collected editions, which also usually contains "The Mudfog Papers" (contributed to *Bentley's Miscellany,* 1837–38); *American Notes* (1842); *Pictures from Italy* (1846); *The Life of Our Lord* (completed 1849, for his children; published 1934); *A Child's History of England* (1852–54); "The Lazy Tour of Two Idle Apprentices" (with Wilkie Collins, contributed to *Household Words* [1857]; often included in the volume entitled *Christmas Stories*); *Reprinted Pieces* (1858; contributed to *Household Words,* 1850–56); *The Uncommercial Traveller* (1861, amplified 1868, 1875; contributed to *All the Year Round,* 1860–69); *Plays and Poems,* ed. by R.H. Shepherd (1885); *Miscellaneous Papers,* ed. by B.W. Matz (1908; the most substantial posthumous collection, mainly essays contributed to *Household Words,* 1850–59; 16 further items, in the volume retitled *Collected Papers,* in *The Nonesuch Dickens,* 1937); *Uncollected Writings from Household Words 1850–1859,* ed. by Harry Stone (1968).

BIBLIOGRAPHY

Bibliographies: K.J. FIELDING, *Charles Dickens* (1953); ADA NISBET, "Charles Dickens," in LIONEL STEVENSON (ed.), *Victorian Fiction: A Guide to Research,* pp. 44–153 (1964, reprinted 1980), a full discussion of materials for Dickens studies and of writings about him in many languages, through 1962; *Victorian Fiction: A Second Guide to Research,* ed. by GEORGE H. FORD, pp. 34–113 (1978), covering 1963–74. See also PHILIP COLLINS, *A Dickens Bibliography* (1970), offprinted from GEORGE WATSON (ed.), *New Cambridge Bibliography of English Literature,* vol. 3, col. 779–850 (1969). REGINALD C. CHURCHILL (comp.), *Bibliography of Dickensian Criticism: 1836–1974* (1975), a selective, partly annotated bibliography.

Most of the manuscripts and proof sheets of the novels are in the Victoria and Albert Museum, London. Other important collections of manuscripts and letters are in Dickens House, London; the British Museum; New York Public Library; Pierpont Morgan Library, New York City; Free Library of Philadelphia; Henry E. Huntington Library and Art Gallery, San Marino, California; the University of Texas Libraries; and Yale University Library. The Dickens Fellowship (Dickens House, London) has branches all over the world and publishes the *Dickensian* (thrice yearly). *Dickens Studies Newsletter* (quarterly) and *Dickens Studies Annual* are published from Carbondale, Illinois, where the Dickens Society is based.

Collected editions: The New Oxford Illustrated Dickens (1947–58); and the *Clarendon* edition, begun in 1966. See also *Speeches,* ed. by K.J. FIELDING (1960); and *Public Readings,* ed. by PHILIP COLLINS (1975).

Letters: The most complete collection, *The Letters of Charles Dickens,* ed. by WALTER DEXTER, 3 vol. (1938), is superseded by *The Letters of Charles Dickens,* ed. by MADELINE HOUSE *et al.,* begun in 1965. See also *The Heart of Charles Dickens, As Revealed in His Letters to Angela Burdett-Coutts,* ed. by EDGAR JOHNSON (1952, reprinted 1976).

Biographies: JOHN FORSTER, *The Life of Charles Dickens,* 3 vol. (1872–74), remains indispensable; though EDGAR JOHNSON, *Charles Dickens: His Tragedy and Triumph,* 2 vol. (1952, reprinted 1965), supersedes it. NORMAN and JEANNE MACKENZIE, *Dickens* (1979), is a popular biography; PHILIP COLLINS (ed.), *Dickens,* 2 vol. (1981), contains interviews with and recollections of people who knew him; FRED KAPLAN, *Dickens and Mesmerism* (1975), relates his interest in hypnotism to concerns expressed in his novels; JOSEPH GOLD, *Charles Dickens: Radical Moralist* (1972), is a discussion of his ethical beliefs.

Criticism: GEORGE R. GISSING, *Charles Dickens: A Critical Study* (1898, reissued 1976); G.K. CHESTERTON, *Charles Dickens* (1903, reprinted 1977); GEORGE ORWELL, "Dickens," in *Critical Essays,* pp. 7–56 (1946); EDMUND WILSON, "Dickens: The Two Scrooges," in *The Wound and the Bow,* pp. 1–104 (1941); HUMPHRY HOUSE, *The Dickens World,* 2nd ed. (1942, reissued 1971), an excellent discussion of Dickens and his age; GEORGE H. FORD, *Dickens and His Readers* (1955, reprinted 1974); JOHN E. BUTT and KATHLEEN TILLOTSON, *Dickens at Work* (1957, reprinted 1982); J. HILLIS MILLER, *Charles Dickens: The World of His Novels* (1958, reissued 1969), a highly influential critical study; PHILIP COLLINS, *Dickens and Crime* (1962); ROBERT GARIS, *The Dickens Theatre* (1965); ANGUS WILSON, *The World of Charles Dickens* (1970); and FRANK R. and Q.D. LEAVIS, *Dickens, the Novelist* (1970, reissued 1979).

Anthologies of Dickens criticism: GEORGE H. FORD and L. LANE (eds.), *The Dickens Critics* (1961, reprinted 1976); STEPHEN WALL (ed.), *Charles Dickens: A Critical Anthology* (1970); and PHILIP COLLINS (ed.), *Dickens, the Critical Heritage* (1971), on his critical reception in 1836–82.

(Ph.C./Ed.)

Digestion and Digestive Systems

In order to sustain themselves, all organisms must obtain nutrients from the environment. Some nutrients serve as raw materials for the synthesis of cellular material; others (*e.g.*, many vitamins) act as regulators of chemical reactions in living cells; and still others, upon oxidation in living cells, yield energy. Not all nutrients, however, are in a form suitable for immediate use by an organism; some must undergo physical and chemical changes before they can serve as energy or cell substance.

This article begins with an overview of digestion and digestive systems, referring to specific organisms to clarify the account; in the latter part, it emphasizes aspects of the human digestive system, its functions, and related diseases and disorders.

For coverage of related topics in the *Macropædia* and *Micropædia*, see the *Propædia*, Part Four, Division II, especially Section 421.

The article is divided into the following sections:

General features of the digestive process 275
- Ingestion 275
- Digestion 276
- Egestion 277

Invertebrate digestive systems 277
- Comparison of unicellular and multicellular organisms 277
- Vacuolar systems
- Channel-network system
- Evolution of cellular specialization 277
  - Saccular systems
  - Tubular systems

Embryology and evolution of the vertebrate digestive system 278
- Embryonic development 278
- Evolutionary development 279

The vertebrate digestive system 280
- Anatomy 280
  - Mouth and oral structures
  - Salivary glands
  - Pharynx
  - Esophagus
  - Stomach
  - Small intestine
  - Large intestine
  - Rectum and anus
  - Liver
  - Biliary tract
  - Pancreas
- Organ function 286
  - Mouth and oral structures
  - Salivary glands
  - Pharynx
  - Esophagus
  - Stomach
  - Small intestine
  - Large intestine
  - Liver
  - Biliary tract
  - Pancreas

Features of the gastrointestinal tract 293
- General features of digestion and absorption 293
- Specific features of digestion and absorption 294
  - Carbohydrates
  - Proteins
  - Fats
  - Fat-soluble vitamins
  - Calcium
  - Magnesium
  - Hematinics
  - Intestinal gas
  - Hormones of the gastrointestinal tract
- The gastrointestinal tract as an organ of immunity 298

Disorders and diseases of the digestive system 299
- Immune-related disorders 299
- Mouth and oral cavity 299
- Salivary glands 300
- Esophagus 300
- Stomach 303
- Duodenum 304
- Small intestine 304
- Large intestine 306
- Liver 308
- Biliary tract 311
- Pancreas 313

Bibliography 313

General features of the digestive process

Through the act of eating, or ingestion, nutrients are taken from the environment. Many nutrient molecules are so large and complex that they must be split into smaller molecules before they can be used by the organism. This process of breaking down food into molecular particles of usable size and content is called digestion. Unusable components are expelled from the organism by a process called egestion, or excretion. Some plants, many microorganisms, and all animals perform these three functions—ingestion, digestion, and egestion (often grouped under the term alimentation)—but, as expected, the details differ considerably from group to group.

Autotrophy and heterotrophy

The problems associated with nutrient intake and processing differ greatly depending on whether the organism is autotrophic or heterotrophic. Autotrophic organisms are those that can manufacture the large energy-rich organic compounds necessary for life from simple inorganic raw materials; consequently, they require only simple nutrients from the environment. By contrast, heterotrophic organisms cannot manufacture complex organic compounds from simple inorganic ones, and so they must obtain preformed organic molecules directly from the environment.

Green plants constitute by far the majority of the Earth's autotrophic organisms. During the process of photosynthesis, they use light energy to synthesize organic materials from carbon dioxide and water. Both compounds can be absorbed easily across the membranes of cells—in a typical land plant, carbon dioxide is absorbed from the air by leaf cells, and water is absorbed from the soil by root cells—and used directly in photosynthesis; *i.e.*, neither of them requires digestion. The only other nutrients needed by most green plants are minerals such as nitrogen, phosphorus, and potassium, which also can be absorbed directly and require no digestion. There are, however, a few green plants (*e.g.*, sundew, Venus's-flytrap, pitcher plant) that supplement their inorganic diet with organic compounds (particularly protein) obtained by trapping and digesting insects and other small animals.

Heterotrophy characterizes all animals, most microorganisms, and plants and plantlike organisms (*e.g.*, fungi) that lack the pigment chlorophyll, which is necessary for photosynthesis. These organisms must ingest organic nutrients—carbohydrates, proteins, and lipids (fats)—and, by digestion, rearrange them into a form suitable for their own particular needs.

INGESTION

As already explained, the nutrients procured by most green plants are small inorganic molecules that can move with relative ease across cell membranes. Heterotrophic organisms such as bacteria and fungi, which require organic nutrients yet lack adaptations for ingesting bulk food, also

rely on direct absorption of small nutrient molecules across cell membranes; molecules of carbohydrates, proteins, or lipids, however, are too large and complex to move easily across cell membranes. Bacteria and fungi circumvent this by secreting digestive enzymes onto the food material; these enzymes catalyze the splitting of the large molecules into smaller units that are then absorbed into the cells. In other words, the bacteria and fungi perform extracellular digestion—digestion outside cells—before ingesting the food. This is often referred to as osmotrophic nutrition.

Ingestion by simple cells

Like bacteria, protozoans are unicellular organisms, but their method of feeding is quite different. They ingest relatively large particles of food and carry out intracellular digestion (digestion inside cells) through a method of feeding called phagotrophic nutrition. To a lesser degree many protozoans also are osmotrophic. Some organisms (*e.g., Amoeba*) put out pseudopodia ("false feet"), which flow around the food particle until it is completely enclosed in a membrane-bounded chamber called a food vacuole; this process (Figure 1A) is called phagocytosis. Other protozoans (*e.g., Paramecium*) pinch off food vacuoles from the end of a prominent oral groove into which food particles are drawn by the beating of numerous small, hairlike projections (cilia). In still other cases of phagotrophic nutrition, tiny particles of food adhere to the membranous surface of the cell, which then folds inward and is pinched off as a vacuole; this process (Figure 1B) is called pinocytosis. The food particles contained in vacuoles formed through phagocytosis or pinocytosis have not entered the cell in the fullest sense until they are digested into molecules able to cross the membrane of the vacuole and become incorporated into the cellular substance.

Most multicellular animals possess some sort of digestive cavity—a chamber opening to the exterior via a mouth—in which digestion takes place. Large particles of food are broken down to units of more manageable size within the cavity before being taken into cells and reassembled (or assimilated) as cellular substance.

(A) Adapted from James G. Hirsch in A.E. Nourse, *The Body*. (B) from W.T. Keeton, *Biological Science*. © 1967 by W.W. Norton & Company, Inc.

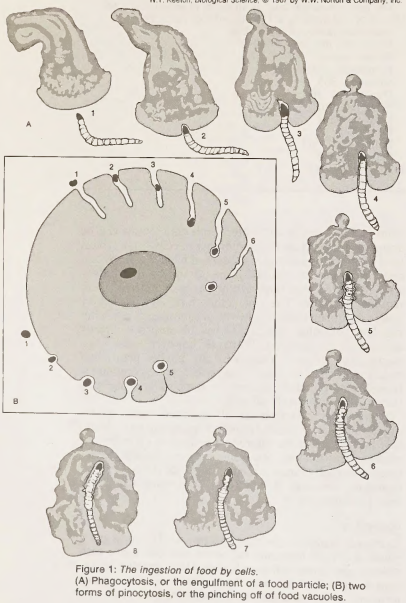


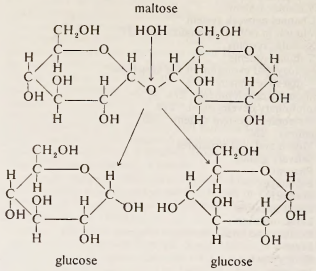
Figure 1: *The ingestion of food by cells.*
(A) Phagocytosis, or the engulfment of a food particle; (B) two forms of pinocytosis, or the pinching off of food vacuoles.

DIGESTION

The enzymatic splitting of large and complex molecules into smaller ones is effective only if the enzyme molecules come into direct contact with the molecules of the material they are to digest. In animals that ingest very large pieces of food, only the molecules at the surface are exposed to the digestive enzymes. Digestion can proceed more efficiently, therefore, if the bulk food is first mechanically broken down, exposing more molecules for digestion. Among the variety of devices that have evolved to perform such mechanical processing of food are the chewing teeth of mammals and the muscular gizzards of birds.

The chemical basis of digestion

The chemical reactions involved in digestion can be clarified by an account of the digestion of maltose sugar. Maltose is, technically, a double sugar, since it is composed of two molecules of the simple sugar glucose bonded together. The digestive enzyme maltase catalyzes a reaction in which a molecule of water is inserted at the point at which the two glucose units are linked, thereby disconnecting them, as illustrated below.



In chemical terms, the maltose has been hydrolyzed. All digestive enzymes act in a similar way and thus are hydrolyzing enzymes.

Many other nutrient molecules are much more complex, being polymers, or long chains of simple component units. Starch, for example, is a carbohydrate, like maltose, but its molecules are composed of thousands of glucose units bonded together. Even so, the digestion of starch is essentially the same as the digestion of maltose: each linkage between adjacent glucose units is hydrolyzed, with the result that the starch molecule is split into thousands of glucose molecules. Protein molecules also are polymers, but their constituent units are amino acids instead of simple sugars. Proteolytic (*i.e.*, protein-digesting) enzymes split the protein chains by hydrolyzing the bonds between adjacent amino acids. Because as many as 20 different kinds of amino acids may act as building blocks for proteins, the complete digestion of a protein into its amino acids requires the concerted action of several different proteolytic enzymes, each capable of hydrolyzing the bonds between particular pairs of amino acids. Fat molecules, too, are composed of smaller building-block units (the alcohol glycerol plus three fatty acid groups); they are hydrolyzed by the enzyme lipase.

Various other classes of compounds are digested by hydrolytic enzymes specific for them. Not all of these enzymes occur in every organism; for example, few animals possess cellulase (cellulose-digesting enzyme), despite the fact that cellulose constitutes much of the total bulk of the food ingested by plant-eating animals. Some nonetheless benefit from the cellulose in their diet because their digestive tracts contain microorganisms (known as symbionts) capable of digesting cellulose; the herbivores absorb some of the products of their symbionts' digestive activity.

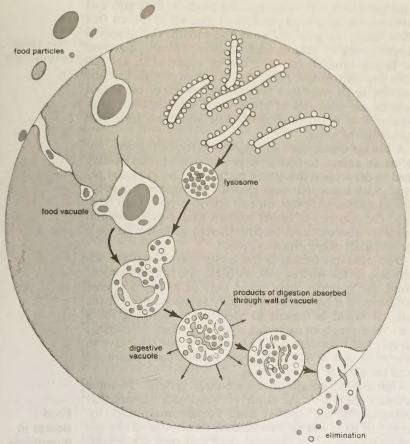


Figure 2: *The role of lysosomes in intracellular digestion.* Digestion takes place when a food vacuole and a lysosome unite, forming a digestive vacuole. The products of digestion are absorbed across the vacuolar membrane, and the indigestible wastes are ultimately expelled to the outside.

Adapted from W.T. Keeton, *Biological Science*, © 1967 by W.W. Norton & Company, Inc.

So far, emphasis has been placed on digestion's role in converting large complex molecules into smaller simpler ones that can move across membranes, thus permitting absorption of food into cells. The same processes occur when substances must be moved from cell to cell within a multicellular organism. Thus green plants, which do not have to digest incoming nutrients, digest stored material, such as starch, before it can be transported from storage organs (tubers, bulbs, corms) to points of utilization, such as growing buds.

EGESTION

The elimination of indigestible matter

Animals that ingest bulk food unavoidably take in some matter that they are incapable of using; for example, since man lacks cellulase, the cellulose he ingests in vegetables and fruits is indigestible. It cannot be absorbed from the digestive tract, and the residue that is not broken down by bacteria must be expelled from the body.

In the case of unicellular organisms that form food vacuoles, the vacuoles eventually fuse with the cell membrane and then rupture, releasing indigestible wastes to the outside (Figure 2). Animals periodically release such waste from their digestive tracts either by regurgitating it through the mouth or by eliminating it as feces through the anus.

Fecal constituents in species with an alimentary canal also include cast-off effete (damaged or worn-out) cells from the living mucous membrane (mucosa) and, in higher animals, bacteria that exist in the intestine in a symbiotic relationship. In the higher animals, the life span of a cell from the mucosal epithelium is four to eight days, and the life span of the specialized cells, such as the acid-secreting parietal cells located in the stomach, is one to three years.

Invertebrate digestive systems

COMPARISON OF UNICELLULAR AND MULTICELLULAR ORGANISMS

Vacuolar systems. Unicellular organisms that ingest food particles via vacuoles rely on intracellular digestion to prepare the nutrients for use. The enzymes that catalyze this digestion, being very potent chemicals capable of breaking down the cell substance itself, are held until needed in special packets, or vesicles, called lysosomes; the membrane of a lysosome is both impermeable to the enzymes and capable of resisting their hydrolytic action. Soon after a food vacuole is formed, a lysosome fuses with it (Figure 2). Food material and digestive enzymes are mixed in the resulting composite vesicle, which is sometimes called a digestive vacuole. This vacuole moves in an orderly fashion through the cell, during which passage the products of digestion are absorbed, leaving the indigestible material, which is eventually expelled.

Vacuolar digestion is not restricted to unicellular organisms. Many multicellular invertebrates partly digest their food extracellularly before phagocytizing the remainder, which is then digested by the process described above.

Channel-network system. The sponges, among the simplest multicellular organisms, have what amounts to diversionary water channels that serve to bring water and food to their component cells. The channels are lined with special cells bearing whiplike structures called flagella that create water currents. A steady flow of water inward through smaller secondary channels and then out the main, or excurrent, canal carries with it bits of food. The lining cells capture the food particles and enclose them in food vacuoles, wherein the matter is digested as in protozoans—by intracellular means.

EVOLUTION OF CELLULAR SPECIALIZATION

Saccular systems. With the evolution of multicellular organisms came a corresponding evolution of cellular specialization, resulting in a division of labour among cells; in this way, certain cells became specialized to perform the function of digestion for the entire organism. Cnidarians, especially hydra, provide a simple example. These radially

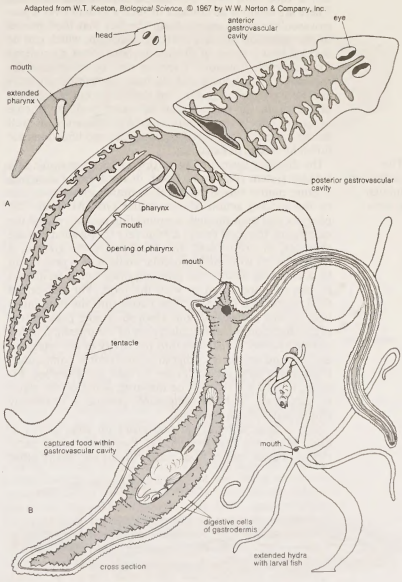


Figure 3: *Saccular systems.* (A) Much-branched gastrovascular cavity and extruded pharynx of a planarian; (B) gastrovascular cavity of a hydra.

symmetrical animals have a saclike body composed of two principal layers of cells (Figure 3B). The cells of the outer layer function as a protective and sensory covering (epithelium); those of the inner layer, or gastrodermis, which lines the central cavity of the body, act as a nutritive epithelium. The central cavity, functioning as a digestive cavity, has only one opening to the outside; the opening acts both as a mouth for ingestion and as an anus for egestion. Such a digestive cavity is called a gastrovascular cavity, because in many animals it has vessel-like branches that convey the contents to all parts of the body.

The gastrovascular cavity

Once prey, captured by a hydra's tentacles, has been passed through the mouth into the gastrovascular cavity, digestive enzymes are secreted into the cavity by the gastrodermal cells, and extracellular digestion begins. In cnidarians, this extracellular digestion is limited largely to partial hydrolysis of proteins. As soon as the food has been partially disintegrated, the gastrodermal cells engulf the fragments by phagocytosis, and digestion is completed intracellularly within food vacuoles.

Many flatworms (phylum Platyhelminthes) also have gastrovascular cavities, even though their bodies are much more complex than those of cnidarians. In planarians, for example, the mouth opens into a tubular chamber called the pharynx, which in turn leads into a branched gastrovascular cavity that ramifies throughout the body (Figure 3A). As in cnidarians, some extracellular digestion occurs in the planarian gastrovascular cavity, with the small food particles then being engulfed by gastrodermal cells and digested intracellularly. The additional process of extracellular digestion frees cnidarians and flatworms from exclusive reliance on intracellular digestion.

Tubular systems. Most animals above the level of cnidarians and flatworms have a complete digestive tract; *i.e.,* a tube with two openings—a mouth and an anus. There are obvious advantages of such a system over a gastrovascular cavity, among them the fact that food moves in one direction through the tubular system, which can be divided into a series of distinct sections, each specialized for a different function. A section may be specialized for mechanical breakdown of bulk food, for temporary storage, for enzymatic digestion, for absorption of the products of digestion, for reabsorption of water, and for storage of wastes. The overall result is greater efficiency, as well as the potential for special evolutionary modifications for different modes of existence.

The earthworm tubular system

The digestive system of an earthworm is an example of a tubular system (Figure 4). Food, in the form of decaying organic matter mixed with soil, is drawn into the mouth by the sucking action of a muscular pharynx. From the pharynx and then through a connecting passage, called the esophagus, the food enters a relatively thin-walled storage chamber, or crop. Next, the food enters the gizzard, a compartment with thick muscular walls, and is ground up by a churning action, the grinding often being facilitated by bits of stone taken in with the food. The pulverized food, suspended in water, then passes into the long intestine, in which digestion and absorption take place. Most of the digestion is extracellular; cells of the intestinal lining secrete hydrolytic enzymes into the cavity of the intestine, and the end products of digestion, the simple compounds from which large molecules are formed, are absorbed. Finally, toward the rear of the intestine, some of the water is reabsorbed, and the indigestible residue is ultimately eliminated through the anus.

Not all large animals eat and grind up large pieces of food. Many are filter feeders; *i.e.,* they strain small particles of organic matter from water. Clams and many other mollusks filter water through tiny pores in their gills and trap microscopic food particles in streams of mucus that flow along the gills and enter the mouth; the mucus is kept moving by beating cilia. In such mollusks, digestion is largely intracellular, as might be expected in animals that eat microscopic food. Current theory holds that the earliest vertebrates were filter feeders. Some of the largest whales are examples of modern-day filter-feeding vertebrates; they strain small planktonic organisms from vast quantities of water.

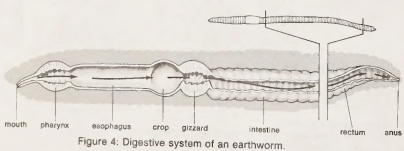


Figure 4: Digestive system of an earthworm.

A storage organ, such as the crop of the earthworm, enables an animal to take in large amounts of food quickly and to draw upon this stored matter over an extended period. Such a discontinuous feeding habit makes it possible for an animal to devote time to activities other than feeding. The majority of higher animals have evolved adaptations for discontinuous feeding, thereby gaining time for a behaviorally more varied existence.

Discontinuous feeding is frequently also of adaptive advantage in the feeding process itself. An animal's proper food, for example, may occur only in widely scattered locations; if it had to eat constantly to maintain itself, the animal would be unable to spend time searching for a new food supply or capturing more prey when the original supply had been depleted. The animal would thus have to live in an area in which there was an essentially unlimited and continuous source of food.

Animal food-storage organs are quite variable. In some animals they take the form of blind sacs (diverticula) branching off the digestive tract. Female mosquitoes, for example, have a large diverticulum that opens off the anterior portion of the digestive tract and runs posteriorly, occupying much of the abdominal cavity. The female mosquito locates a suitable animal, pierces its skin, and sucks blood until the diverticulum is filled. One large meal may suffice for the entire process of locating a site and laying her eggs—a matter of four or five days.

Food storage in mosquitoes

Embryology and evolution of the vertebrate digestive system

EMBRYONIC DEVELOPMENT

In amphioxus, an invertebrate member of the Chordata (the phylum to which all vertebrates belong), early divisions of the fertilized egg cell give rise to an embryo that is hollow and nearly spherical. Then an invagination (infolding) of cells at the vegetal (yolk) pole of the embryo converts the initially single-layered embryo into a two-layered one, a process called gastrulation. The new inner layer of cells, called endoderm (sometimes entoderm), surrounds a cavity, the archenteron, which has an opening to the exterior at the point at which invagination occurred; this opening is called the blastopore. The archenteron eventually becomes the cavity of the digestive tract, and the blastopore becomes the anus; the mouth arises as a new opening. In some invertebrates the reverse is true: the blastopore becomes the mouth, and the anus is the new opening.

The early stages of embryonic development in most vertebrates are not as simple as in amphioxus, largely because the egg cells contain much yolk or, in mammals, undergo specialized changes preparatory to implantation in the uterus. Thus, gastrulation is seldom a simple involution at the vegetal pole, and the blastopore, if indeed a "pore" appears at all, usually becomes overgrown with cells. Nevertheless, in all vertebrate embryos, an endodermal-lined cavity arises by some process that may be regarded as analogous to gastrulation in amphioxus, and this cavity develops into the digestive tract. Ordinarily, endoderm lines the yolk sac; forms a tube, called the foregut, that pushes forward into the head; and forms a second tube, the hindgut, that pushes into the posterior part of the embryonic body. Eventually, the surface tissue (ectoderm) of the embryo forms a small anterior invagination, the stomodeum, that meets the end of the foregut, and a similar posterior invagination, the proctodeum, that meets the end of the hindgut. Rupture of the tissues separating the stomodeum from the foregut and the proctodeum from the hindgut forms a tract with two openings to the exterior.

Embryonic derivations

It is apparent from the above description that short sections at both the anterior and the posterior ends of the digestive tract are of ectodermal origin. These correspond roughly to the oral cavity and to the anal canal, respectively. All the rest of the digestive tract, from the pharynx through the large intestine, is of endodermal origin. However, only the lining of the digestive tract is endodermal; the walls contain layers of muscle and connective tissue, which are of middle layer (mesodermal) origin. The endodermal lining gives rise by outpocketing to numerous organs, including the thyroid gland, gills or lungs, thymus, liver, pancreas, and urinary bladder.

EVOLUTIONARY DEVELOPMENT

In amphioxus, the digestive tract consists of only three components: the oral cavity, the pharynx, and a tubular postpharyngeal gut without subdivisions. The same condition holds in the most primitive living vertebrates, the cyclostomes (lampreys and hagfishes). In higher vertebrates, however, the postpharyngeal gut is almost always subdivided into a series of regions both anatomically and functionally distinct. The most common is the esophagus–stomach–small intestine–large intestine–rectum (or cloaca) sequence.

The oral cavity and pharynx vary considerably among the vertebrate classes. The variation correlates with the evolutionary changes in the respiratory system that accompanied the rise of terrestrial forms from aquatic ancestors. In most modern-day bony fishes, the nares (corresponding to a mammal's nostrils) function only as entrances to the olfactory organs, there being no connection between them and the mouth, as occurs in mammals. The structure called the palate, which in mammals separates the nasal and oral cavities, does not exist in fishes. Respiratory water is taken directly into the mouth and then forced back into the pharynx, where it flows across gills located in a series of slits leading from the pharynx to the exterior. Thus, the pharynx, with its gills, is an extremely important chamber in these animals.

The terrestrial vertebrates, which must extract oxygen from air instead of from water, evolved a second major function for the nares that they inherited from their piscine ancestors. While retaining a smell function, these openings became the principal entrance of air for breathing. In amphibians—the earliest land vertebrates—air enters the external nares (nasal openings) and then passes through the internal nares, which are evolutionarily newer openings, into the front of the oral cavity, whence it moves into the pharynx and then into the trachea. There being no palate, no separate nasal cavity exists in these animals; both the oral cavity and the pharynx are common passages for the digestive and respiratory systems.

In most reptiles and birds, a pair of longitudinal folds in the roof of the oral cavity forms a passage that leads air from the internal nares to the pharynx. Complete separation of nasal and oral cavities by a palate, however, is found only in crocodilians and in mammals. In mammals, the bony, hard palate is supplemented posteriorly by a thick, membranous, soft palate.

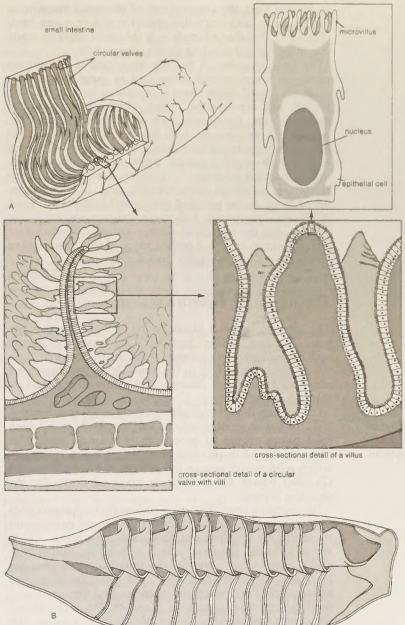


Figure 6: Structural modifications for increasing the surface areas of the small intestine.

Adapted from *Human Biology* by G.A. Baitsell; Copyright 1950; used with permission of McGraw-Hill Book Co.

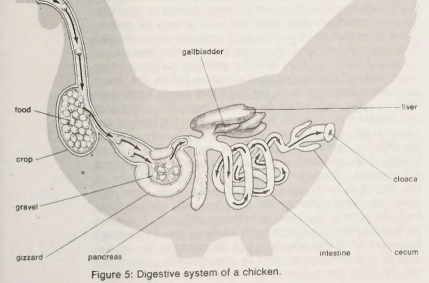


Figure 5: Digestive system of a chicken.

Changes in the esophagus

In the evolution of terrestrial vertebrates the pharynx has lost the gas-exchanging gills and has become a short passage linking the mouth to the esophagus and the trachea. The esophagus has elongated to join up with the stomach, which now lies within the abdomen.

Most vertebrates above the level of the cyclostomes have a stomach, though of various shapes and sizes (the exceptions are the chimaeras, lungfishes, and a few bony fishes). The length of the intestine varies greatly among vertebrates, and a number of devices have evolved that increase the area over which absorption of digestive products can occur. Increasing length alone permits longer contact between the product of digestion and the mucosa. Other features of advantage include the lining of the intestine, which is thrown into numerous folds and ridges; the small, fingerlike outgrowths, called villi, that cover the entire surface of the mucosa; and the individual epithelial cells that cover the folds and villi and have a border of countless, closely packed, cylindrical projections called microvilli (Figure 6A).

Other vertebrates show other adaptations for increasing the absorptive surface area of the small intestine. For instance, it is not unusual for special blind sacs, called ceca,

to branch from the anterior end of the small intestine in certain fishes and from the posterior end in many birds (Figure 5). Another adaptation is the spiral valve of many primitive fishes, including sharks (Figure 6B).

The final chamber of the digestive tract is a common cloaca in elasmobranch fishes and in lungfishes, but in most ray-finned fishes there is a rectum instead; *i.e.*, the urinary and reproductive tubes, which do not join the digestive tube, have their own separate opening to the exterior. In this regard, then, the modern-day ray-finned fishes are more specialized than amphibians, reptiles, and birds, which retain a cloaca, presumably inherited from a primitive fish ancestor. A cloaca is also retained in the egg-laying mammals (monotremes), and, in a much reduced form, in the pouched mammals (marsupials). Even in placental mammals, a short-lived cloaca appears in the embryo, but the urogenital ducts eventually develop their own openings, and, as a consequence of this, mammalian adults have a rectum rather than a cloaca.

(W.T.Ke./W.S.)

The vertebrate digestive system

The account below is based on the human digestive system. The first section deals with anatomy, the second with physiology. Both proceed through the anatomical divisions from the mouth to the anus. Other vertebrates are mentioned when they illustrate some important departure through specialization.

The human digestive system

The human digestive system consists of (1) the digestive tract, or the series of structures and organs through which food passes during its processing into forms absorbable into the bloodstream and also the structures through which solid wastes pass in the process of elimination, and (2) other organs that contribute juices necessary for the digestive process.

The digestive tract (Figure 7) begins at the lips and ends at the anus. It consists of the mouth, or oral cavity, with its teeth, for grinding the food, and its tongue, which serves to knead the food, mix it with saliva, and start it on its way to the stomach; the throat, or pharynx; the esophagus, or gullet; the stomach; the small intestine, consisting of the duodenum, the jejunum, and the ileum; and the large intestine, consisting of the cecum, a closed-end sac connecting with the ileum, the ascending colon, the transverse colon, the descending colon, and the sigmoid colon, which terminates in the rectum. Glands contributing digestive juices include the salivary glands, the gastric glands in the stomach lining, the pancreas, and the liver and its adjuncts—the gallbladder and bile ducts.

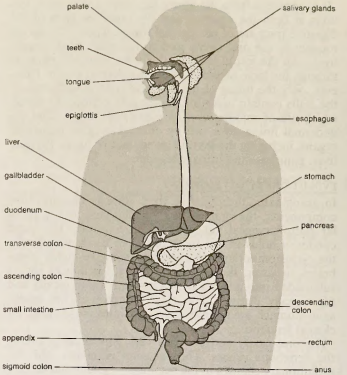


Figure 7: The human digestive system as seen from the front.

ANATOMY

Mouth and oral structures. *The lips.* The outer surface of the lips is covered with skin, the inner with mucous membrane. The mucosa is rich in mucus-secreting glands, which together with salivary mucus ensure adequate lubrication for the purposes of speech and mastication.

The cheeks. The cheeks, the sides of the mouth, are continuous with the lips and have a similar structure. A distinct fat pad is found in the subcutaneous tissue (the tissue beneath the skin) of the cheek; this pad is especially large in infants and is known as the sucking pad. On the inner surface of each cheek, opposite the second upper molar tooth (the second grinder tooth from the end), is the slight elevation that marks the opening of the parotid duct, leading from the parotid salivary gland, which is located in front of the ear. Just behind this gland are four to five mucus-secreting glands, the ducts of which open opposite the last molar tooth.

The roof of the mouth. The roof of the mouth is concave and is formed by the hard and soft palate. The hard palate is formed by the horizontal portions of the two palatine bones and the palatine portions of the maxillae, or upper jaws. The hard palate is covered by a thick, somewhat pale mucous membrane that is continuous with that of the gums and is bound to the upper jaw and palate bones by firm fibrous tissue. The soft palate is continuous with the hard palate in front and posteriorly has a free margin. The uvula, which varies greatly in size and shape, is a projection of the soft palate and hangs free from its rear margin. The soft palate is composed of a strong, thin, fibrous sheet, the palatine aponeurosis, and the glossopalatine and pharyngopalatine muscles.

The hard and soft palates

The floor of the mouth. The floor of the mouth can be seen only when the tongue is raised. In the midline is a prominent fold (frenulum linguae) like that which binds each lip to the gums, and on each side of this is a slight elevation called a sublingual papilla, onto the summit of which the ducts of the submaxillary salivary glands open. Running outward and backward from this is a ridge (the plica sublingualis) that marks the upper edge of the sublingual (under the tongue) salivary gland and onto which most of the ducts of that gland open.

The gums. The gums consist of mucous membranes connected by thick fibrous tissue to the membrane surrounding the bones of the jaw. Around the base of the crown (exposed portion) of each tooth the gum membrane rises to form a little collar. The gum tissues are rich in blood vessels, receiving branches from the alveolar arteries; these vessels, called alveolar from their relationship to the alveoli dentales, or tooth sockets, also supply the teeth and the spongy bone of the upper and lower jaws, in which the teeth are lodged. The veins and lymphatics of the gums correspond essentially to the arteries.

Specialized functions of teeth

The teeth. The teeth are hard white structures found in the mouth of humans and many other animals and usually are used for mastication. The teeth of different vertebrate species are sometimes specialized. The teeth of snakes, for example, are very thin and sharp and usually curve backward; they function in capturing prey but not in chewing, for snakes swallow their food whole. The teeth of carnivorous mammals, such as cats and dogs, are more pointed than those of primates, including humans; the canines are long, and the premolars lack flat grinding surfaces, being more adapted to cutting and shearing (often the more posterior molars are lost). On the other hand, herbivores such as cows and horses have very large, flat premolars and molars with complex ridges and cusps; the canines are often totally absent. Sharp pointed teeth, poorly adapted for chewing, generally characterize meat eaters such as snakes, dogs, and cats; and broad, flat teeth, well adapted for chewing, characterize vegetarians. The differences in the shapes of teeth are functional adaptations. Few animals can digest cellulose, yet the plant cells used as food by herbivores are enclosed in cellulose cell walls that must be broken down before the cell contents can be exposed to the action of digestive enzymes. By contrast, the animal cells in meat are not encased in nondigestible matter and can be acted upon directly by digestive enzymes. Conse-

quently, chewing is not so essential for carnivores as for herbivores. Dogs gulp their food; cows and horses spend much time chewing. Humans, who are omnivores (eaters of plants and animal tissue), have teeth that belong, functionally and structurally, somewhere between the extremes of specialization attained by the teeth of carnivores and herbivores. The members of one class of vertebrates, the birds, have no true teeth. Mechanical breakdown of food is accomplished in birds in a muscular gizzard (Figure 5) located behind the stomach; in the gizzard, hard food is ground with grit formed of ingested rocks and pebbles.

Each tooth consists of a crown and one or more roots. The crown is the functional part that is visible above the gum. The root is the unseen portion that supports and fastens the tooth in the jawbone. The shapes of the crowns and the roots vary in the different parts of the mouth and from one animal to another. Humans normally have two sets of teeth during their lifetime. The first set is acquired gradually between the ages of six months and two years. As the jaws grow and expand, these teeth are replaced one by one by the teeth of the secondary set. The first set is known as the milk, deciduous, or primary dentition. The teeth on one side of the jaw are essentially a mirror image of those located on the opposite side. The upper teeth differ from the lower and are complementary to them. There are five deciduous teeth and eight permanent teeth in each quarter of the mouth, making a total of 32 permanent teeth to succeed the 20 deciduous ones.

Functions of the tongue

The tongue. The tongue, a muscular organ located on the floor of the mouth, is an extremely mobile structure in humans and an important accessory organ in such motor functions as speech, chewing, and swallowing. In conjunction with the cheeks, it is able to guide and maintain food between the upper and lower teeth until mastication is completed. The tongue's motility aids in creating a negative pressure within the oral cavity, thus enabling mammals to suckle.

The mucous membrane that covers the tongue varies greatly. Especially important as a peripheral sense organ, it contains groups of specialized epithelial cells, known as taste buds, that carry stimuli from the oral cavity to the central nervous system. Furthermore, the tongue's glands produce some of the saliva necessary for swallowing.

The mammalian tongue consists of a mass of interwoven, striated (striped) muscles covered with mucous membrane and interspersed with glands and a variable amount of fat. By its extrinsic muscles, the tongue is attached to the lower jaw, the hyoid bone (a U-shaped bone between the lower jaw and the larynx), the skull, the soft palate, and the pharynx. It is bound to the floor of the mouth and to the epiglottis (a plate of cartilage that serves as a lid for the larynx) by folds of its mucous membrane.

Salivary glands. Besides the many minute glands that secrete saliva, there are three major pairs of salivary glands: the parotid, the submaxillary, and the sublingual glands.

Size, form, and location

The parotid glands, the largest of the pairs, are located at the side of the face, below and in front of each ear, in back touching the mastoid bone (which is behind the ear) and the sternomastoid muscle, the prominent muscle of each side of the neck, and in front shaped around the ascending portion of the bone of the lower jaw. The parotid glands are enclosed in sheaths that limit the extent of their swelling when inflamed, as in mumps. The submaxillary glands lie near the inner side of the lower jawbone, not far in front of the sternomastoid muscle. They are rounded in shape. The sublingual glands lie directly under the mucous membrane covering the floor of the mouth beneath the tongue.

The salivary glands are of the type called racemose, from the Latin *racemosus* ("full of clusters"), because of the cluster-like arrangement of their secreting cells in rounded sacs, called acini, attached to freely branching systems of ducts. The walls of the acini surround a small central cavity known as an alveolus. In the walls are pyramidal secreting cells and some flat, star-shaped cells called myoepithelial, or basket, cells. The latter cells are thought to contract, like the similar myoepithelial cells of the breast, which by their contraction expel milk from the milk ducts.

The secreting cells may be of the serous or the mucous type. The latter type secretes mucin, the chief constituent of mucus; the former, a watery fluid containing an enzyme, amylase, which is also known as ptyalin. The secreting cells of the parotid glands are of the serous type; those of the submaxillary glands, of both serous and mucous types, with the serous cells outnumbering the mucous cells by four to one. The acini of the sublingual glands are composed primarily of mucous cells.

Types of secreting cells

The parasympathetic nerve supply regulates secretion by the acinar cells and causes the blood vessels to dilate. Functions regulated by the sympathetic nerves include secretion by the acinar cells, constriction of blood vessels, and, presumably, contraction of the myoepithelial cells.

Pharynx. The pharynx, or throat, is the passageway leading from the mouth and nose to the esophagus and larynx. The pharynx permits the passage of swallowed solids and liquids into the esophagus, or gullet, and conducts air to and from the trachea, or windpipe, during respiration. The pharynx also connects on either side with the cavity of the middle ear by way of the Eustachian tube and provides for equalization of air pressure on the eardrum membrane, which separates the cavity of the middle ear from the external ear canal.

The pharynx has roughly the form of a funnel flattened from front to back. It is attached to the surrounding structures but is loose enough in organization to permit gliding of the pharyngeal wall against them, in the movements of swallowing.

Regional anatomy

Three main divisions of the pharynx are distinguished: the oral pharynx, the nasal pharynx, and the laryngeal pharynx. The latter two are airways, whereas the oral pharynx is shared by both the respiratory and alimentary (digestive) tracts. On either side of the opening between the mouth cavity and the oral pharynx is a tonsil called a palatine tonsil because of its proximity to the palate. Each palatine tonsil is between two vertical folds of mucous membrane (behind the glossopalatine arch and in front of the pharyngopalatine arch). The glossopalatine arches are located on the sides of the pharynx. The nasal pharynx, above, is separated from the oral pharynx by the soft palate. The laryngeal pharynx and the lower part of the oral pharynx are hidden by the bulging root of the tongue. An important feature of this obscured region is the epiglottis, or laryngeal flap, which acts as a deflector between the laryngeal pharynx and the lowermost oral pharynx and closes over the larynx during the act of swallowing.

The pharyngeal muscles are concerned in the mechanics of swallowing. The principal muscles of the pharynx are the three pharyngeal constrictors, which overlap each other slightly, from below upward, and form the primary musculature of the side and rear pharyngeal walls.

A series of lymphatic glands known as the tonsils encircle the oropharyngeal junction. Besides the palatine tonsils, there is one pair on the roof of the nasopharynx that, when grossly swollen, which is often during childhood, occlude the airway there, and then are known as adenoids. They are part of the body's immune-defense system.

Esophagus. The esophagus, which extends from the pharynx to the stomach, is about 25 centimetres (10 inches) in length; the width varies from one and one-half to two centimetres. The esophagus contains the four typical layers of the alimentary canal—mucosa (or mucous membrane), submucosa, muscularis, and tunica adventitia. The mucosa is made up of stratified squamous epithelium containing numerous mucous glands. The submucosa is a thick, loose fibrous layer connecting the mucosa to the muscularis. The mucosa and submucosa form long longitudinal folds so that a cross section of the esophagus opening would be star-shaped. The muscularis is composed of an inner layer in which the fibres are circular and an outer layer of longitudinal fibres. In fact both muscle groups are wound around and along the alimentary tract, but the inner one has a very tight spiral so that the windings are virtually circular, and the outer one has a very slowly unwinding spiral that is virtually longitudinal.

Layers of the esophagus

The outer layer of the esophagus, the tunica adventitia, is composed of loose fibrous tissue that connects the esophagus with neighbouring structures. Except during the act of swallowing, the esophagus is normally empty, and its

lumen, or channel, is essentially blocked by the longitudinal folds of the mucosal and the submucosal layers.

For the upper third of the esophagus the muscle is striated (capable of being contracted by volition). The middle third has a mixture of striated and smooth (involuntary) muscle, and the lower third only of smooth muscle. The esophagus has two sphincters. (Sphincters are circular muscles that act like drawstrings in closing channels.) The upper esophageal sphincter is located at the level of the cricoid cartilage (a single ringlike cartilage forming the lower part of the larynx wall). This sphincter is called the cricopharyngeus muscle. The lower esophageal sphincter encircles the three to four centimetres of the esophagus that pass through the diaphragmatic hiatus and is thus located partly above and partly below the diaphragm. Both sphincters normally remain closed except during the act of swallowing.

The lower esophageal sphincter

Stomach. The stomach serves as a reservoir and receives ingested food and liquids from the esophagus and retains them for admixing with the gastric juice in order that digestion can begin. It is located in the left upper part of the abdomen immediately below the diaphragm. In front of the stomach are the liver, part of the diaphragm, and the anterior abdominal wall. Behind it are the pancreas, the left kidney, the left adrenal, the spleen, and the colon. When the stomach is empty, it contracts, and the transverse colon ascends to occupy the vacated space.

The size, shape, and position of the stomach vary extremely and depend upon the extent of its contents as well as upon the tension in the muscles of its walls. The organ is more or less concave on its right side, convex on its left. The concave border is called the lesser curvature; the convex border, on the left and below, the greater curvature. The opening from the esophagus into the stomach is the cardia, while the outlet from the stomach into the duodenum is the pylorus.

The parts of the stomach

The various parts of the stomach may be summarized as follows. The uppermost part, located above the entrance of the esophagus, is the fundus; it frequently contains a gas bubble, especially after a meal. The cardia is that portion of the stomach surrounding the opening from the esophagus. The largest part of the stomach is the body; it serves primarily as a reservoir for ingested food and liquids. The antrum, the lowermost part of the stomach, is somewhat funnel-shaped, with its wide end joining the lower part of the body of the stomach and its narrow end connecting with the pyloric canal, which empties into the duodenum. The pyloric portion (antrum plus pyloric canal) of the stomach tends to curve to the right and slightly upward and backward and thus gives the stomach its J-shaped appearance. The pylorus, the narrowest portion of the stomach, is approximately two centimetres in diameter; it is surrounded by thick loops of smooth muscle.

When the stomach is empty, its mucosal lining is thrown into numerous longitudinal folds, known as rugae; these tend to disappear when the stomach is distended.

The surface of the gastric (stomach) mucosa is always covered by a layer of thick tenacious mucus that is secreted by the columnar cells of the surface. Beneath the surface epithelium various types of gastric glands are located that, in different parts of the stomach, vary in structure and in composition of their secretion. Thus, each area of the gastric mucosa is characterized by its glandular structure. When the gastric mucus is removed from the surface epithelium, small pits, called foveolae gastricae, may be observed with a hand magnifying glass. There are approximately 90 to 100 gastric pits per square millimetre (58,000 to 65,000 per square inch) of surface epithelium. Into each gastric pit from three to seven individual gastric glands empty their secretions.

The gastric mucosa contains five different types of cells. In addition to the tall columnar surface epithelial cells mentioned above, there are three common cell types found in the various gastric glands. (1) Mucoid cells—cells that secrete mucus—are common to all types of gastric glands. Mucoid cells are the main cell type found in the gastric glands in the cardiac and pyloric areas of the stomach. The necks of the glands in the body and fundic parts of the stomach are lined with mucoid cells. (2) Other cells, called zymogenic, or chief, cells, are located predominantly in the gastric glands in the body and fundic portions of the stomach. These cells secrete pepsinogen, from which the enzyme pepsin is formed. (3) Cells called parietal, or oxyntic, cells, in the glands of the body and fundic portions of the stomach, secrete hydrochloric acid, most of the water found in gastric juice, and a protein called intrinsic factor (see below *Organ function*). (4) Endocrine cells, called enterochromaffin-like cells because of their staining characteristics, are scattered throughout the body of the stomach. (5) Other endocrine cells throughout the antrum secrete the acid-stimulating hormone gastrin.

The gastric glands

Beneath the gastric mucosa is a thin layer of smooth muscle fibres called the muscularis mucosa, and below this, in turn, is loose connective tissue, the submucosa, which attaches the gastric mucosa to the muscles in the walls of the stomach.

The stomach muscles

The muscles of the gastric wall are arranged in three layers, or coats. The innermost layer of smooth muscle of the gastric wall, called the oblique muscular layer, is strongest in the region of the gastric fundus and progressively weaker as it approaches the pylorus.

The middle, or circular muscular layer, the strongest of the three muscular layers, completely covers the stomach. The circular fibres of this coat are best developed in the lower portion of the stomach, particularly over the antrum and the pylorus.

The external coat, called the longitudinal muscle layer, is formed from the longitudinal muscle coat of the esophagus. The longitudinal muscle fibres are divided at the cardia into two broad strips. The one on the right, the stronger, spreads out to cover the lesser curvature and the adjacent posterior and anterior walls of the stomach. The longitudinal fibres on the left sweep from the esophagus over the dome of the fundus of the stomach to cover the greater curvature and continue on to the pylorus, where they join the longitudinal muscle fibres coming down over the lesser curvature.

At the pyloric end of the stomach the middle layer of muscle, the circular layer, becomes greatly thickened to form the pyloric sphincter. This muscular ring is slightly separated from the circular muscle of the duodenum by connective tissue. The outer muscle layer of the stomach, the longitudinal layer, continues on into the duodenum, forming the longitudinal muscle of the small bowel.

Many branches of the celiac trunk bring arterial blood to the stomach. The celiac trunk is a short, wide artery that branches from the abdominal portion of the aorta, the main vessel conveying arterial blood from the heart to the systemic circulation. Blood from the stomach is returned to the venous system into the portal vein, which carries the blood to the liver.

Nerve supply

The nerve supply to the stomach is provided by both the parasympathetic and sympathetic divisions of the autonomic nervous system. The parasympathetic nerve fibres are carried in the vagus, or 10th cranial, nerves. As the vagus nerves pass through the opening in the diaphragm together with the esophagus, branches of the right vagus nerve spread over the posterior part of the stomach, while the left vagus supplies the anterior part. Sympathetic branches from a nerve network called the celiac, or solar, plexus accompany the arteries of the stomach into the muscular wall.

Small intestine. The small intestine, which is 670–760 centimetres in length, is the longest part of the digestive tract of humans. It begins at the pylorus, the juncture with the stomach, and ends at the ileocecal valve, the juncture with the colon. The parts of the small intestine are the duodenum, jejunum, and ileum.

The three parts of the small intestine. The duodenum is 23–28 centimetres long and forms a horseshoe, or C-shaped, curve that encircles the head of the pancreas. Unlike the rest of the intestine, it is retroperitoneal (that is, it is behind the peritoneum, the membrane lining the abdominal wall). Its first part, known as the duodenal bulb, is the widest part of the small intestine. It is horizontal, passing backward and to the right from the pylorus, and lies somewhat behind the wide end of the gallbladder. The second part runs vertically downward in front of the

hilum of the right kidney (the point of entrance or exit for blood vessels, nerves, and the ureters): it is into this part of the duodenum, through the papilla of Vater, that the pancreatic juice and bile flow. The third part runs horizontally to the left in front of the aorta and the inferior vena cava (the principal channel for return to the heart of venous blood from the lower part of the body and the legs), while the fourth part ascends to the left side of the second lumbar vertebra (at the level of the small of the back), then bends sharply downward and forward to join the second part of the small intestine, the jejunum. An acute angle, called the duodenojejunal flexure, is formed by the suspension of this part of the small intestine by the ligament of Treitz.

The jejunum and ileum

The jejunum forms the upper two-fifths of the rest of the small intestine; it, like the ileum, has numerous convolutions and is attached to the posterior abdominal wall by mesentery, a fold of serous—clear-fluid-secreting—membrane.

The ileum is the remaining three-fifths of the small intestine, though there is no absolute point at which the jejunum ends and the ileum begins. In broad terms, the jejunum occupies the upper and left part of the abdomen below the subcostal plane (*i.e.*, from a plane just above the floating ribs), while the ileum is located in the lower and right part. At its termination the ileum opens into the large intestine.

The small intestine mucosa. Although the small intestine is only three to four centimetres in diameter and approximately seven metres (23 feet) in length, it has been estimated that the total absorptive area of the human small intestine is approximately 4,500 square metres (5,400 square yards). This enormous absorptive surface is provided by the unique structure of the mucosa, which is arranged in concentric folds that have the appearance of transverse ridges. These folds are approximately five to six centimetres in length and about 3.2 millimetres thick. They are known as plicae circulares. These folds are present throughout the small bowel except in the first portion, or bulb, of the duodenum, which is usually flat and smooth, except for a few longitudinal folds. The plicae circulares are largest in the lower part of the duodenum and in the upper part of the jejunum. They become smaller and finally disappear in the lower part of the ileum. It has been estimated that the small intestine of humans contains approximately 800 plicae circulares and that they increase the surface area of the lining of the small bowel by five to eight times the outer surface of the small intestine.

Another feature of the small bowel mucosa that greatly multiplies its surface area is that of the tiny projections called mucosal villi. These mucosal villi usually vary from 0.5 to one millimetre in height. Their diameters vary from approximately one-eighth to one-third their height. The villi are covered by a single layer of tall columnar cells. Goblet cells, so called because of their rough resemblances to empty goblets after they have discharged their contents, are also found scattered among the surface cells. The goblet cells are a source of mucin, the chief constituent of mucus.

Micro-structure

At the base of the mucosal villi are depressions called intestinal glands, or Lieberkühn's glands. The cells that line these glands continue up and over the surface of the villi. In the bottom of the glands, the epithelial cells are filled with alpha granules, or eosinophilic granules, so called because they take up the rose-coloured stain eosin. These cells have been termed the cells of Paneth. Though they may contain lysozyme, an enzyme toxic to bacteria, and immunoglobins, their precise function is uncertain.

There are three other cell types in these glands: undifferentiated cells that have the potential for undergoing changes for the purpose of replacing losses of any cell type, the goblet cells mentioned above, and endocrine cells, which are described below. The main functions of the undifferentiated cells in these glands are cell renewal and secretion. The surface epithelium coating the villi, on the other hand, is concerned with digestion and absorption. The epithelial cells covering the villi move progressively upward, maturing into one cell form or another. Those coating the villi have an average life of 72 hours in humans and most larger animals, before becoming exhausted and being cast off.

Anatomy of villi

In the duodenum the villi are closely packed and large and frequently are leaflike in shape. In the jejunum the individual villus measures between 350 and 600 microns in height (there are about 25,000 microns in an inch) and has a diameter of 110 to 135 microns. The appearance and shape of the villi vary in different levels of the small intestine. The inner structure of the individual villus consists of loose connective tissue containing a rich network of blood vessels, a central lacteal, or channel for lymph, smooth muscle fibres, and scattered cells of various types. The smooth muscle cells surround the central lacteal and provide for the pumping action required to initiate the flow of lymph out of the villus.

A small central arteriole (minute artery) branches at the tip of the villus to form a capillary network (the capillaries are the smallest of the blood vessels); the capillaries, in turn, empty into a collecting venule that runs to the bottom of the villus.

The vascular channels in the villi have a thin endothelial cell lining with tiny diaphragm-covered pores that have greater permeability to large molecules than the nonporous parts. Villus structure varies in different ethnic groups, but this may only reflect differences in diet and parasitism, especially with the worms that are endemic in Eastern and tropical countries.

The microvilli

The most remarkable feature of the small bowel mucosa is the rough surface of the epithelial cells, the cells covering the villi. This surface, called a brush border, consists of individual microvilli. The microvilli are approximately one-tenth micron in diameter and one micron in height; each epithelial cell may have as many as 1,000 microvilli. The microvilli play an important role in absorption of intestinal contents by enlarging the absorbing surface approximately 25 times and containing a number of enzymes.

Beneath the mucosa of the small intestine, as beneath that of the stomach, are the muscularis mucosae and the submucosa. The submucosa consists of loose connective tissue and contains many blood vessels and lymphatics. In the submucosa of the duodenum are located Brunner's glands, composed of acini (round sacs) and tubules that are twisting and have multiple branching. Brunner's glands empty into the base of Lieberkühn's glands in the duodenum. Their exact function is not known, but they do secrete a scanty clear fluid that contains mucus, bicarbonate, and a relatively weak proteolytic (protein-splitting) enzyme. In the submucosa of the jejunum, solitary nodules (lumps) of lymphatic tissue are located. There is more lymphatic tissue in the ileum, in aggregates of nodules known as Peyer's patches.

The arrangement of the muscular coats of the small intestine is uniform throughout the length of the intestine. There is an inner, circular layer that is thicker than the outer, longitudinal layer. The outermost covering of the small intestine is the peritoneum, a single layer of flattened epithelial cells.

The blood supply of the small intestine is from the superior mesenteric artery (a branch of the abdominal aorta) and from the superior pancreaticoduodenal artery (a branch of the hepatic artery). These vessels run between layers of the mesentery and give off large branches that form a row or series of connecting arches from which branches enter the wall of the small bowel. The blood from the intestine is returned by means of the superior mesenteric vein, which, with the splenic vein, forms the portal vein, which drains into the liver.

The small intestine has both sympathetic and parasympathetic innervation. The vagus nerves provide the parasympathetic innervation. The sympathetic is provided by branches from the superior mesenteric plexus, a nerve network close under the solar plexus, which follow the blood vessels into the small intestine and finally terminate in the two plexuses, or networks—Auerbach's, located between the circular and longitudinal muscle coats, and Meissner's, which is located in the submucosa. Numerous fibrils, both adrenergic (sympathetic) and cholinergic (parasympathetic) connect the two plexuses.

Large intestine. The large intestine serves as a reservoir

Anatomy of the colon

for the liquids emptied into it, through the ileocecal valve, from the small intestine. It has a much larger diameter than the small intestine. The large intestine, or colon, may be divided into the cecum, ascending colon, transverse colon, descending colon, and sigmoid colon. The primary function of the colon is to absorb water and electrolytes (substances, such as salts, that in solution take on an electrical charge) from the ileal contents and to store fecal material until it can be evacuated by defecation.

The cecum, the first part of the large intestine, is a sac with a closed end. It occupies the right iliac fossa, the hollow of the inner side of the ilium (the upper part of the hipbone). Guarding the opening of the ileum into the cecum is the ileocecal valve. Most textbooks of anatomy describe the ileocecal valve as consisting of two folds or flaps of mucous membrane on the cecal side and above and below the ileal opening. The ileocecal junction does have this appearance commonly after death. During life, however, the ileocecal junction appears much different in that the terminal portion of the ileum doubles into the cecum; from the cecal side the ileocecal valve greatly resembles the cervix, the projection of the uterus into the vagina. The circular muscle fibres of the ileum and those of the cecum combine to form the circular sphincter muscle of the ileocecal valve.

Regional anatomy

The ascending colon extends up from the cecum at the level of the ileocecal valve to the bend in the colon called the hepatic flexure, which is located beneath and behind the right lobe of the liver; behind, it is in contact with the rear abdominal wall and the right kidney. The ascending colon is covered by peritoneum except on its posterior surface.

The transverse colon is variable in position, depending largely on the distention of the stomach, but usually is located in the subcostal plane; that is, at the level of the 10th rib. On the left side of the abdomen it ascends to the bend called the splenic flexure, which may make an indentation in the spleen. The transverse colon is bound to the diaphragm opposite the 11th rib by a fold of peritoneum.

The descending colon passes down and in front of the left kidney and the left side of the posterior abdominal wall to the iliac crest, the upper border of the hipbone. The descending colon is more likely than the ascending colon to be surrounded by peritoneum.

The sigmoid colon is commonly divided into iliac and pelvic parts. The iliac colon stretches from the crest of the ilium, or upper border of the hipbone, to the inner border of the psoas muscle, which lies in the left iliac fossa. Like the descending colon, the iliac colon is usually covered by peritoneum. The pelvic colon lies in the true pelvis (lower part of the pelvis) and forms one or two loops, reaching across to the right side of the pelvis and then bending back and, at the midline, turning sharply downward to the point where it becomes the rectum.

The layers that make up the wall of the colon are similar in some respects to those of the small bowel; there are distinct differences, however. The external aspect of the colon differs markedly from that of the small bowel because of features known as the haustra, taeniae, and appendices epiploicae.

The haustra, bulges or sacculations, are formed by constricting circular furrows of varying depths. The three taeniae are long, narrow bands of longitudinal muscle fibres, about one centimetre in width, that are approximately equally spaced around the circumference of the colon. Between the thick bands of the taeniae there is a thin coating of longitudinal muscle fibres.

The appendices epiploicae are collections of fatty tissue beneath the covering membrane. On the ascending and descending colon they are usually found in two rows, whereas on the transverse colon they form one row.

Microstructure

The mucous membrane of the colon has a characteristic structure. It lacks the villi and the folds known as plicae circulares characteristic of the small intestine. It contains many solitary lymphatic nodules but no Peyer's patches. The surface epithelium is columnar, and there are many goblet cells. Characteristic of the colonic mucosa are deep tubular pits, increasing in depth toward the rectum.

The arterial blood supply to the large intestine is supplied by branches of the superior and inferior mesenteric arteries (both of which are branches of the abdominal aorta) and the hypogastric branch of the internal iliac (which supplies blood to the pelvic walls and viscera, the genital organs, the buttocks, and the inside of the thighs). The vessels form a continuous row of arches from which vessels arise to enter the large intestine. Venous blood is drained from the colon from branches that form venous arches similar to those of the arteries. These eventually drain into the superior and inferior mesenteric veins, which ultimately join with the splenic vein to form the portal vein.

The innervation of the large intestine is similar to that of the small intestine.

Rectum and anus. The rectum, which is a continuation of the sigmoid colon, begins in front of the midsacrum (the sacrum is the triangular bone near the base of the spine and between the two hipbones). It ends in a dilated portion called the rectal ampulla, which in front is in contact with the rear surface of the prostate in the male and with the posterior vaginal wall in the female. Posteriorly, the rectal ampulla is in front of the tip of the coccyx (the small bone at the very base of the spine).

Regional anatomy

At the end of the pelvic colon, the mesocolon, the fold of peritoneum that attaches the colon to the rear wall of the abdomen and pelvis, ceases, and the rectum is then covered by peritoneum only at its sides and in front; lower down, the rectum gradually loses the covering on its sides, until only the front is covered. At about the junction of the middle and lower thirds of the rectum, about 7.5 centimetres from the anus, the anterior peritoneal covering is also folded back onto the bladder and the prostate or the vagina.

Near the termination of the sigmoid colon and the beginning of the rectum, the colonic taeniae spread out to form a wide external longitudinal muscle coat. At the lower end of the rectum muscle fibres of the longitudinal and circular coats tend to intermix. The internal circular muscle coat terminates in the thick rounded internal anal sphincter muscle. The smooth muscle fibres of the external longitudinal muscle coat of the rectum terminate by interweaving with striated muscle fibres of the levator ani, a broad muscle that forms the floor of the pelvis. A second sphincter, the external anal sphincter, is composed of striated muscle and is divided into three parts known as the subcutaneous, superficial, and deep external sphincters. Thus, the internal sphincter is composed of smooth muscle and is innervated by the autonomic nervous system, while the external sphincters are of striated muscle and have somatic (voluntary) innervation provided by nerves called the pudendal nerves.

The mucosal lining of the rectum is similar to that of the sigmoid colon but becomes thicker and better supplied with blood vessels, particularly in the lower rectum. In the rectal ampulla are two to three large crescent-like folds known as rectal valves. These folds, or valves, are caused by an invagination, or infolding, of the circular muscle and submucosa. The columnar epithelium of the rectal mucosa changes to the stratified squamous (scalelike) type in the lower rectum a few centimetres above the pectinate line, which is the junction between squamous mucous membrane of the lower rectum and the skin lining the lower portion of the anal canal.

The rectal valves

Arterial blood is supplied by branches from the inferior mesenteric artery and the right and left internal iliac arteries. Venous drainage from the anal canal and rectum is provided by a rich network of veins called the internal and external hemorrhoidal veins. (N.C.H./W.S.)

Liver. The liver is not only the largest gland in the human body, but it is also the most complex in function. In humans, the liver lies under the lower right rib cage and occupies much of the upper right quadrant of the abdomen, with a portion extending into the upper left quadrant (Figure 8). The liver weighs from 1.2 to 1.6 kilograms (2.6 to 3.5 pounds) and is somewhat larger in men than in women. Its greatest horizontal measurement ranges from 20 to 22 centimetres; vertically, it extends 15 to 18 centimetres and in thickness it ranges from 10 to 13 centimetres. The liver is reddish brown, and its outline, when viewed from the front, resembles an asymmetrical

Lobes of the liver

rhomboid whose lateral vertical dimension is about three times as long as the medial side. In humans and other primates, the liver is divided into two unequal lobes: a large right lobe and a smaller left lobe. The left lobe is separated on its anterior (frontal) surface by the dense falciform (sickle-shaped) ligament that connects the liver to the undersurface of the diaphragm. On the inferior surface of the liver, the right and left lobes are separated by a groove containing the teres ligament, which runs to the umbilicus. Two small lobes, the caudate and the quadrate, occupy a portion of the inferior surface of the right lobe. The entire liver, except for a small portion that abuts the right leaf of the diaphragm, is enveloped in a capsule of tissue that is continuous with the parietal peritoneum, which lines the abdominopelvic walls and diaphragm.

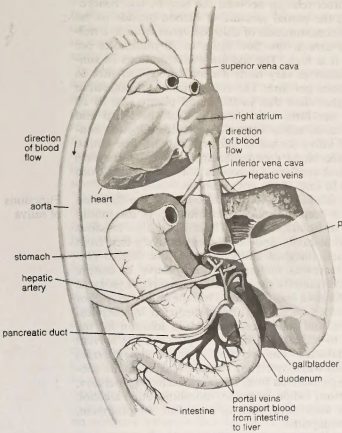


Figure 8: The liver and adjacent organs, as seen from behind.

The major blood vessels enter the liver on its inferior surface in a centrally placed groove called the porta hepatis, which anatomically separates the quadrate and caudate lobes. The liver has two sources of blood supply: fully oxygenated blood from the hepatic artery, which is a major branch of the celiac axis (the main artery that crosses the abdomen) after its emergence from the abdominal aorta; and partially oxygenated blood from the large portal vein, which in turn receives all venous blood from the spleen, pancreas, gallbladder, lower esophagus, and the remainder of the gastrointestinal tract, including the stomach, small intestine, colon, and upper portion of the rectum. The portal vein is formed by the juncture of the splenic vein with the superior mesenteric vein. At the porta hepatis, the portal vein divides into two large branches, each going to one of the major lobes of the liver. The hepatic duct, which also exits at the porta, is the final pathway for a network of smaller bile ductules interspersed throughout the liver that serve to carry newly formed bile from liver cells to the small intestine.

Micro-scopic anatomy

The microscopic anatomy of the liver reveals a uniform lobular structure throughout. Cords made up of numerous rectangular liver cells, or hepatocytes, are found to radiate from central veins, or terminal hepatic venules, toward a peripheral boundary defined by scant rims of connective tissue that establish the lobular border. A lobule measures about one millimetre in diameter. The liver cords are one-cell thick and are separated from one another on several surfaces by spaces called sinusoids, or hepatic capillaries. Sinusoids are lined by thin endothelial cells that have holes in their body (cytoplasm) through which fingerlike projections (microvilli) of the hepatocytes extend, allowing direct accessibility of the hepatocyte to the bloodstream in the sinusoids. The other major cell of the liver, the Kupffer cell, adheres to the wall of the sinusoid at intervals and projects into its lumen. It functions as a phagocyte (a cell that engulfs and destroys foreign material or other cells). Small spaces (Disse's spaces), probably for the transport of lymph, are present in places between the hepatocyte and the sinusoidal endothelium. On apposing surfaces, the hepatocytes are bound to one another by dense, tight junctions. These are perforated by small channels, called canaliculi, that are the terminal outposts of the biliary system, receiving bile from the hepatocyte. They eventually join with other canaliculi, forming progressively larger bile ducts that eventually emerge from the porta as the hepatic duct.

At the periphery of each hepatic lobule are found several areas containing three vessels. These portal areas contain terminal branches of the portal vein and the hepatic artery, conveying blood to the sinusoids, and a small bile duct that carries bile from the hepatic canaliculi. A single hepatic lobule may receive arterial and portal venous blood from and discharge bile into several portal areas.

Biliary tract. The extrahepatic biliary tract commences with the appearance of two large ducts, the right and left hepatic ducts, at the porta hepatis. Just below the porta, these two one- to two-centimetre ducts join to form the hepatic duct, which proceeds for another two to three centimetres and is joined by the cystic duct, leading from the gallbladder. The resulting common bile duct progresses downward through the head of the pancreas. There it is usually joined by the main pancreatic duct of Wirsung at a slightly dilated area called the ampulla of Vater, which lies in the wall of the inner curve of the descending duodenum, and terminates in the lumen of the duodenum at a two- to three-centimetre elevation called the papilla of Vater. The common bile duct averages about 10 centimetres in length, and flow of bile from its lower end into the intestine is controlled by the muscular action of the sphincter of Oddi located in the papilla of Vater. Total daily flow of bile into the intestine ranges from 600 to 800 millilitres. The cystic duct varies from two to three centimetres in length and terminates in the gallbladder, a saccular structure with a capacity of about 50 millilitres. Throughout its length, the cystic duct is lined by a spiral mucosal elevation, called the valvula spiralis (valve of Heister). Normally, the gallbladder lies partially embedded on the undersurface of the right lobe of the liver.

Common bile duct

Pancreas. The human pancreas weighs between 90 and 120 grams. It is a long, narrow gland that is situated transversely across the upper abdomen in the space behind the posterior peritoneum. Its somewhat bulbous head lies against the inner curve of the C-shaped loop formed by the curve of the second portion of the duodenum. The body and tail of the pancreas rise upward and leftward from the head, lying behind the stomach and the spleen. The midportion of the body lies against the vertebral column, the abdominal aorta, and the inferior vena cava.

Dual glandular role of the pancreas

The pancreas is both an exocrine (ductal) and endocrine (ductless) gland. The exocrine acinar tissue produces important digestive enzyme precursors that are transmitted into the small intestine, while the islets of endocrine tissue (islets of Langerhans) produce at least two hormones (insulin and glucagon) that are important in the regulation of carbohydrate metabolism and others (vasoactive intestinal polypeptide and somatostatin) that are pivotal elements in the control of intestinal secretion and motility. Acinar and ductal tissue arise from multiple sites along the primordial epithelial buds, and islet cell tissue arises from other sites. These two separate tissues mix with one another during early fetal life, but each retains its own distinctive character, acinar tissue being linked to the ductal system, and islet tissue having no ductal connections. Distinctly separate islet and acinar tissue are detectable microscopically by the 19th week of intrauterine life.

Histologically, individual acinar cells have the shape of a truncated pyramid, arranged in groups around a central ductal lumen. These central ducts empty into progressively larger intercalated and collecting ducts that eventually join

the duct of Wirsung. This main pancreatic duct enters the ampulla of Vater, where, in about 80 percent of instances, it is joined by the common bile duct. Occasionally the junction with the common bile duct is proximal to the ampulla, and in a few cases the duct of Wirsung and the common bile duct join the duodenum separately. There are at least three types of islet cells, designated alpha (or A), beta (or B), and delta (or D), which constitute about 2 percent of the total pancreatic mass. Islet cells are about 20 to 35 percent alpha, 60 to 75 percent beta, and 5 percent delta. Alpha cell granules contain only glucagon, whose release leads to the breakdown of glycogen in the liver and elevation of the level of blood glucose, while beta cell granules contain insulin, whose effects are the opposite of glucagon. (H.J.Dw.)

ORGAN FUNCTION

Mouth and oral structures. Little digestion of food takes place in the mouth; however, food is prepared in the mouth for transport through the upper alimentary canal, thus aiding the digestive processes that take place in the stomach and small intestine.

Mechanical activities

Mastication, or chewing, is the first mechanical process to which food is subjected. Movements of the lower jaw in chewing are brought about by the muscles of mastication (the masseter, the temporal, the medial and lateral pterygoids, and the buccinator). The sensitivity of the peridontal membrane that surrounds and supports the teeth, rather than the power of the muscles of mastication, limits the force of the bite.

Mastication is not essential for adequate digestion and nutrition. Chewing does aid digestion, however, by reducing food to small particles and mixing it with the saliva secreted by the salivary glands. The saliva lubricates and moistens dry food, while chewing distributes the saliva throughout the food mass. The movement of the tongue against the hard palate and the cheeks helps to form a rounded mass, or bolus, of food.

Salivary glands. The oral cavity has other functions in addition to those associated with mastication. It is there that food is tasted and mixed with saliva secreted by several sets of salivary glands. The saliva dissolves some of the food and acts as a lubricant, facilitating passage through the subsequent portions of the digestive tract. The saliva of some mammals, including humans, also contains a starch-digesting enzyme called amylase (ptyalin), which initiates the process of enzymatic hydrolysis; it splits starch (a polysaccharide containing many sugar molecules bound in a continuous chain) into molecules of the double sugar maltose. Many carnivores, such as dogs and cats, have no amylase in their saliva; hence their natural diet contains very little starch.

Control of salivary secretion

The salivary glands are controlled by the two divisions of the autonomic nervous system (sympathetic and parasympathetic), and it is generally held that this innervation is exclusively responsible for regulation of the glands' secretory activity. No hormone appears to be involved. Normally secretion of saliva is constant, regardless of the presence of food in the mouth. When something touches the gums, the tongue, or some region of the mouth lining, or when chewing occurs, the amount of saliva secreted increases. The stimulating substance need not be food—dry sand in the mouth or even moving the jaws and tongue when the mouth is empty are effective in increasing the salivary flow. This coupling of direct stimulation to the oral mucosa with increased salivation is known as the unconditioned salivary reflex. When an individual learns that a particular sight, sound, smell, or other stimulus is regularly associated with food, that stimulus alone may suffice to stimulate increased salivary flow. This response is known as the conditioned salivary reflex.

A variety of drugs are capable of increasing or decreasing salivary flow. Two types of drugs that increase the flow of saliva are sympathomimetic agents, or drugs that produce effects similar to those elicited by the sympathetic nervous system (*e.g.*, epinephrine [adrenaline], norepinephrine [noradrenaline], and amphetamine), and parasympathomimetic agents, or drugs that mimic the effects of the parasympathetic nervous system (*e.g.*, acetylcholine and pilocarpine). Among the drugs that decrease salivary flow are antagonists of epinephrine and norepinephrine (*e.g.*, ergotamine and dibenzylchlorethamine) and antagonists of acetylcholine (*e.g.*, atropine and scopolamine).

Chemistry

The composition of saliva varies, but its principal components are water, inorganic ions similar to those commonly found in the blood plasma, and a number of organic constituents. The amount of saliva secreted by a human in 24 hours usually amounts to one to 1.5 litres. Although saliva is slightly acidic, the bicarbonates and phosphates contained within it serve as buffers and maintain the pH, or hydrogen ion concentration, of saliva relatively constant under ordinary conditions.

The concentrations of bicarbonate, chloride, and sodium in saliva are directly related to the rate of flow. There is also a direct relation between the bicarbonate concentration and the partial pressure of carbon dioxide in the blood. The concentration of chloride varies from five millimoles per litre at low flow rates, to 70 millimoles per litre when it is high. The sodium concentrations in similar circumstances vary from five millimoles per litre to 100 millimoles per litre. The concentration of potassium is often higher than that in the blood plasma, up to 20 millimoles per litre, which accounts for the sharp and metallic taste of saliva when flow is brisk. Organic constituents of saliva consist of salivary proteins, free amino acids, specific blood group substances, and the enzymes lysozyme and amylase. Glucose is normally absent from saliva even in individuals with diabetes.

Functions of saliva

The main functions of saliva are to initiate the digestion of starch, to keep the mucous membranes of the mouth moist, and to facilitate speech. The constant flow of saliva keeps the oral cavity and teeth comparatively free from food residues, sloughed epithelial cells, and foreign particles. By removing material that may serve as culture media, saliva inhibits the growth of bacteria. Lysozyme serves a protective function, for it has the ability to lyse, or dissolve, certain bacteria. Taste is mediated by chemical mechanisms, and substances must be in solution for the taste buds to be stimulated. Saliva provides the solvent for food materials. The secretion of saliva also provides a mechanism whereby certain organic and inorganic substances can be excreted from the body, including mercury, lead, potassium iodide, bromide, morphine, ethyl alcohol, and certain antibiotics, such as penicillin, streptomycin, and chlortetracycline.

Although saliva is not essential to life, its absence results in a number of inconveniences, including dryness of the oral mucous membrane, poor oral hygiene because of bacterial overgrowth, a greatly diminished sense of taste, and difficulties with speech.

Pharynx. The first stage of deglutition, or swallowing, consists of passage of the bolus into the pharynx and is initiated voluntarily. The front part of the tongue is retracted and depressed, mastication ceases, respiration is inhibited reflexly, and the back portion of the tongue is elevated and retracted against the hard palate. This action, produced by the strong muscles of the tongue, forces the bolus from the mouth into the pharynx. Pressures as great as 100 centimetres of water have been recorded in the posterior part of the oral cavity during swallowing.

At the same time, the larynx moves upward and forward under the base of the tongue. The superior pharyngeal constrictor muscles contract, initiating a rapid pharyngeal peristaltic, or squeezing, contraction that moves down the pharynx, propelling the bolus in front of it. The walls and structures of the lower pharynx are elevated to engulf the oncoming bolus. The epiglottis, a lidlike covering that protects the entrance to the larynx, diverts the bolus to the pharynx. The cricopharyngeal muscle, or upper esophageal sphincter, which has kept the esophagus closed until this point, relaxes as the bolus approaches and allows it to enter the upper esophagus. The pharyngeal peristaltic contraction continues into the esophagus and becomes the primary esophageal peristaltic contraction.

The swallowing reflex is coordinated so well that food normally takes only one path, namely, that into the esophagus. Return of the bolus into the mouth is prevented by the position of the tongue against the hard palate. Entry

into the nasal pharynx is prevented by the elevation of the soft palate against the posterior pharyngeal wall. Entry into the larynx is prevented because the larynx is drawn under the base of the tongue and because the epiglottis diverts material away from the laryngeal opening.

Esophagus. Transport through the esophagus is accomplished by the primary esophageal peristaltic contractions that originated in the pharynx. The primary esophageal peristaltic contraction is produced by an advancing peristaltic wave that creates a pressure gradient and sweeps the bolus ahead of it. Transport of material through the esophagus requires approximately 10 seconds. When the bolus arrives at the junction with the stomach, the lower esophageal sphincter relaxes and the bolus enters the stomach. If the bolus is too large, or if the peristaltic contraction is too weak, the bolus may become arrested in the middle or lower esophagus. When this occurs, secondary peristaltic contractions originate around the bolus in response to the local distension of the esophageal wall and propel the bolus into the stomach.

Esophageal function

In some vertebrates the esophagus is not merely a tubular connection between the pharynx and the stomach, but rather may serve as a storage reservoir, or an ancillary digestive organ. In many birds, for example, an expanded region of the esophagus anterior to the stomach forms a thin-walled crop (Figure 5), which is the bird's principal organ for the temporary storage of food. Some birds also use the crop to carry food to their young. Ruminant mammals, such as the cow, are often said to have four "stomachs." Actually, the first three of these chambers (rumen, reticulum, and omasum) are thought to be derived from the esophagus. Vast numbers of bacteria and protozoans live in the rumen and reticulum. When food enters these chambers, the microbes begin to digest and ferment it, breaking down not only protein, starch, and fats, but cellulose as well. The larger, coarser material is periodically regurgitated as the cud, and after further chewing, the cud is reswallowed. Slowly the products of microbial action, and some of the microbes themselves, move into the cow's true stomach and intestine, and further digestion and absorption take place. Since the cow, like other mammals, has no cellulose-digesting enzymes of its own, it relies upon the digestive activity of these symbiotic microbes in its digestive tract. Much of the cellulose in the cow's herbivorous diet, which otherwise would have no nutritive value, is thereby made available to the cow.

In humans, the tone of the lower esophageal sphincter is maintained at all times, except in response to a descending contraction wave, at which point it relaxes momentarily to allow the eructation of gas (belching) or vomiting. The sphincter has an important role, therefore, in protecting the esophagus from the reflux of gastric contents with changes in body position or with alterations of intragastric pressure. In animals, gastroesophageal reflux is prevented in different ways. For example, in the three-toed sloth, which spends much of its time hanging upside down, there are three large septa, or folds, in the upper end of the stomach, which are aligned so that when the upper stomach is distended the septa come into juxtaposition and completely shut off the cavity of the stomach from the esophagus.

Lower esophageal sphincter

When a liquid bolus is swallowed, its transport through the esophagus depends somewhat on the position of the body and the effects of gravity. When swallowed in a horizontal or head-down position, liquids are handled in the same manner as solids, with the liquid bolus moving immediately ahead of the advancing peristaltic contraction. (The high pressures and strong contractions of the esophageal peristaltic wave make it possible for animals with very long necks, such as the giraffe, to transport liquids through the esophagus to a height of many feet.) When the body is in the upright position, liquids enter the esophagus and fall to the lower end and await the arrival of the peristaltic contraction and the opening of the lower esophageal sphincter (eight to 10 seconds), before being emptied into the stomach.

Stomach. The uppermost third of the stomach, or fundus, adapts to the varying volume of ingested food by relaxing its muscular wall, holding the food while it undergoes the first stages of digestion. The middle third, or body, of the stomach contains a mucosal lining that has glands which house the chief cells and the parietal cells. The chief cells produce the proteolytic (protein-digesting) enzymes, the pepsins. The parietal cells secrete hydrogen ions into a branching system of tiny canals (canaliculi) within the cell, where they combine with chloride to form hydrochloric acid. The parietal cells also secrete intrinsic factor, a protein that is concerned with the liberation of vitamin B_{12}. Endocrine cells in the mucosal glands of the lower third, or antrum, of the stomach, produce the hormone gastrin. When gastrin is secreted in response to a meal or other stimuli, it enters the bloodstream and is carried in the circulation to the mucosa of the body of the stomach. There it binds to receptor sites on the outer membrane of the parietal cells. The gastrin-receptor complex that is formed triggers the metabolic steps within the cell that lead to the production and secretion of hydrogen ions and the formation of hydrochloric acid.

The stomach serves as a temporary reservoir for ingested food and liquids. Its size, shape, and capacity vary from one person to another and also within the same person. The stomach is capable of dilating to accommodate more than one litre of food or liquids without increasing intraluminal pressure within the stomach. This receptive relaxation of the upper part of the stomach to accommodate a meal is partly due to a neural reflex that is triggered when hydrochloric acid comes into contact with the mucosa of the antrum, possibly through the release of the hormone known as vasoactive intestinal peptide (see below *Small intestine*).

The stomach mixes the food with the gastric juices and renders the ingested meal more soluble, preparing it for further digestion in the small intestine. Factors that influence the digestive activity of the stomach include its mobility, the rate of emptying of the gastric contents into the small intestine, and the secretion of hydrochloric acid and enzymes.

Mobility. Three types of motor activity of the human stomach have been observed. The first is a small contraction wave of the stomach wall that originates in the upper part of the stomach and slowly moves down over the organ toward the pyloric sphincter (the opening into the small intestine at the distal end of the stomach). This type of contraction produces a slight indentation of the stomach wall and is thought to serve the purpose of mixing the gastric contents. The second type of motor activity is also a contracting wave, but it is peristaltic in nature. The contraction originates in the upper part of the stomach as well and is slowly propagated over the organ toward the pyloric sphincter. This type of gastric contraction produces a deep indentation in the wall of the stomach. As the peristaltic wave approaches the antrum, the indentation completely obstructs the lumen, thus compartmentalizing it. The contracting wave then moves over the antrum, propelling the material ahead of it through the pyloric sphincter into the duodenum. This type of contraction serves as a pumping mechanism for emptying the contents of the gastric antrum through the pyloric sphincter. The third type of gastric motor activity is best described as a tonic, or sustained, contraction of all of the muscles of the organ. It decreases the size of the lumen of the stomach, as all parts of the gastric wall seem to contract simultaneously. It is this type of activity that accounts for the stomach's ability to accommodate itself to varying volumes of gastric content. The tonic contraction is independent of the other two types of contractions; however, mixing contractions and peristaltic contractions normally are superimposed upon the tonic contraction.

Types of motor activity

Both the mixing and the peristaltic contractions of the human stomach occur at a constant rate of three per minute when recorded from the gastric antrum. This rate is now recognized as the basic rhythm, although some drugs are capable of abolishing both types of contractions or of stimulating the strength of contractions.

In the higher vertebrates and in humans, a wave of peristalsis sweeps along the lower half of the stomach and along the entire intestine to the proximal colon at two-hour intervals after meals. These peristaltic waves can be

halted by eating and can be induced by an intravenous infusion of a hormone, motilin.

Gastric emptying. The distension of the body of the stomach by food activates a neural reflex that initiates the activity of the muscle of the antrum. Retrograde waves frequently sweep from the pyloric sphincter to the antrum and up to its junction with the body of the stomach, resulting in a to-and-fro movement of the gastric contents that has a mixing and crushing effect. As the food is broken down, small particles flow through the pyloric sphincter, which opens momentarily as a peristaltic wave descends through the antrum toward it. This permits "sampling" of the gastric contents by the duodenum (the uppermost division of the small intestine). Larger particles are swept back for further grinding. When receptors in the duodenal bulb (the area of attachment between the duodenum and the stomach) have a fluidity and a hydrogen ion concentration at a certain level the duodenal bulb and the second part of the duodenum relax, allowing emptying of the stomach to start. The vagus nerve of the central nervous system has an important role in the control of emptying, but studies on paraplegia caused by injury to the spinal cord indicate that the sympathetic division of the autonomic nervous system is also involved.

Emptying process

Liquids empty according to the pressure gradient between the stomach and duodenum. During a duodenal contraction, the pressure in the duodenal bulb rises higher than that in the antrum, but the pylorus then acts like most other alimentary sphincters and prevents reflux by shutting. In pyloroplasty, an operation in which the encircling muscle fibres of the sphincter are cut, there is little effect upon the rate of gastric emptying. Solids empty when the masses are sufficiently reduced in size and nearly soluble. Several of the peptide hormones of the alimentary tract have an effect on intragastric pressure and gastric movements, but their role in physiological circumstances is unclear.

Gastric secretion. The gastric mucosa that lines the stomach of humans secretes 1.2 to 1.5 litres of gastric juice per day. This juice is highly acidic because of its hydrochloric acid content, and it is rich in enzymes. Gastric juice renders food particles soluble, initiates digestion (particularly of proteins), and converts the gastric contents to a semiliquid mass, called chyme, thus preparing it for the small intestine and further digestion. The composition of the gastric juice varies according to the stimulus. It is a mixture of water, hydrochloric acid, electrolytes (sodium, potassium, calcium, phosphate, sulphate, and bicarbonate), and organic substances (mucus, pepsins, and protein).

Hydrochloric acid is produced by the parietal cells when one or more types of receptors on the outer membrane of the parietal cell are bound to one of three substances (histamine, gastrin, or acetylcholine) that have a particular molecular structure specific for a certain type of receptor. Histamine, probably released from mast cells that are present near the parietal cells, is a response to food entering the stomach. Gastrin, released from the endocrine cells in the antrum, is a response to lowered acidity of the gastric contents when food enters the stomach. Acetylcholine, released when stimuli descend the vagus nerve, is a response to the aesthetic and gustatory aspects of feeding and to the physical effects of chewing and swallowing. Histamine release may have a central role as the final common pathway by which the parietal cell is activated, the other stimulants working by first activating the mast cells.

As the gastric contents become less acidic, gastrin is released and travels in the blood to receptor sites on the exterior of the parietal cells of the stomach. In these cells, an energy-consuming reaction moderated by the presence of ATPase bound to the membrane gives off hydrogen ions (protons). Intracellular hydrogen ions are exchanged for extracellular potassium ions, which become vehicles for transporting intracellular chloride ions to the canaliculi, where they form hydrochloric acid (HCl) with free hydrogen ions already extruded from the cell. Inhibition of this proton pump by a drug, omeprazole, stops the secretion of acid and is a promising treatment for peptic ulcer. The acid that is collected in the canaliculi within the parietal cell drains into the lumen of the gland and then passes through to the stomach. The chief cells secrete the zymogen, or proenzyme (precursor), pepsinogen, which is converted to the active enzyme, pepsin, by hydrochloric acid. This conversion is achieved by inserting water into the precursor molecules. Some other enzymes, including lipase and carbonic anhydrase, can be found, but their origin is uncertain.

Production of acid

Gastric mucus is a glycoprotein that serves two purposes: the lubrication of food masses, facilitating movement within the stomach, and the formation of a protective layer over the lining epithelium of the stomach cavity. This protective layer is a defense mechanism the stomach has against being digested by its own protein-lyzing enzymes, and it is facilitated by the secretion of bicarbonate into the surface layer from the underlying mucosa. The physical character of mucus—its viscosity, resistance to penetration by hydrogen ions, and the thickness of the layer over the mucosa—varies, although the mechanisms of this are unknown. Nonsteroidal anti-inflammatory drugs, including aspirin and those used in the treatment of arthritis, degrade mucus and reduce its protective properties. When these agents damage the mucosa, however, bicarbonate is secreted in increasing amounts, so that the result will depend on the balance of these opposing forces.

The acidity, or hydrogen ion concentration, in the "unstirred" layer of mucus on the surface of the mucosa is a gradient that is measured at pH7 (neutral) at the area immediately adjacent to the epithelium and becomes more acidic (pH2) at the luminal level. Pepsin and acid secreted from the glands dissolve mucus glycoprotein in the gastric lumen, but this is replenished continually with viscous undegraded mucus from the mucosa.

Another element in the protection of the stomach walls from digestive juices is the membrane on the surface of the epithelial cells bordering the lumen of the stomach: this membrane is rich in lipoproteins, which are resistant to attack by acid. Most invertebrates do not have proteolytic enzymes active in an acid milieu, whereas most vertebrates have some enzymes activated by acid and others activated by alkali.

Derived from dietary lipids in response to various chemical and hormonal stimuli, the prostaglandins are hormonelike substances that are present in virtually all animal tissues and body fluids. The many different prostaglandins are designated by combinations of letters and numbers according to the chemical structure. Their synthesis in the body reflects the major fatty acids in the diet. In humans, arachidonic acid is the precursor of most prostaglandins. Prostaglandins are involved in the contraction and dilatation of blood vessels, the aggregation of platelets (clotting), and the contraction and relaxation of the smooth muscle of the gastrointestinal tract. Prostaglandins also inhibit the secretion of hydrochloric acid by the stomach in response to food, histamine, and gastrin. On a molecular level, prostaglandin PGE_2 is a thousand times as potent as H_2-receptor antagonists in inhibiting acid secretion. Prostaglandins are able to protect the mucosa of the alimentary tract from injury by various insults, including boiling water, alcohol, aspirin, bile acids, hypertonic saline, and stress. This protection is not related to their ability to influence acid secretion. Prostaglandins increase the secretion of mucus and bicarbonate from the mucosa, and they stimulate the migration of cells to the surface for repair and replacement of the mucosal lining.

Prostaglandins

Vitamin B_{12}-binding glycoproteins, present in gastric juice, are designated IF and IF_R according to the rate at which they move during electrophoresis. They are secreted from the pepsinogen-producing chief cells in some vertebrates and from the acid-producing parietal cells in others, including humans. IF_R binds vitamin B_{12} more strongly than does IF. (Vitamin B_{12} is essential to the maturation of red blood cells and to the health of certain cells in the central and peripheral nervous systems.) Because only a fraction of the total IF and IF_R that is produced is necessary, the number of parietal cells must be markedly diminished before a deficiency state appears. In a rare disorder, IF secretion is absent, although acid production is normal.

There are two varieties of protein-splitting proenzymes,

Pepsinogen

known as pepsinogen I and pepsinogen II. Both are produced in the mucous and chief cells in the glands of the body of the stomach, but the mucous glands that are distributed elsewhere in the stomach produce only pepsinogen II. The ability of the pepsins to break down protein (proteolysis) is restricted by the necessity for an acidic background with a pH between 1.8 and 3.5. Those stimuli that excite gastric acid secretion, in particular, vagal nerve stimulation, also promote the secretion of the pepsinogens. Pepsinogen I is useful as a genetic marker that signals the inheritance of gastric diseases. When acid secretion is supranormal, so, as a rule, is the secretion of pepsinogen I.

Phases of gastric secretion. The process of gastric secretion can be conveniently divided into three phases (cephalic, gastric, and intestinal). These phases depend upon the primary mechanisms that cause the gastric mucosa to secrete gastric juice.

The cephalic phase of gastric secretion occurs in response to stimuli received by the senses; that is, taste, smell, sight, and sound. This phase of gastric secretion is entirely reflex in origin and is mediated by the vagus (10th cranial) nerves. The gastric juice that is secreted in response to vagal stimulation, either directly by electrical impulses or indirectly by stimuli received through the special senses, is rich in enzymes and is highly acidic. Ivan Petrovich Pavlov, the Russian physiologist, originally demonstrated this method of gastric secretion in a now famous experiment in dogs.

The gastric phase is mediated by the vagus nerves and by the release of gastrin. The acidity of the gastric contents after a meal is buffered by food proteins so that overall it remains around pH3 for approximately 90 minutes. Acid continues to be secreted during the gastric phase in response to distension and to the peptides and amino acids that are liberated from food protein as digestion proceeds. The chemical action of free amino acids and peptides excites the liberation of gastrin from the antrum into the circulation. Thus, there are mechanical, chemical, and hormonal factors contributing to the gastric secretory response to eating. This phase continues until the food leaves the stomach.

The secretion of acid itself is an important inhibitor of gastrin release. If the pH of the antral contents falls below 2.5, gastrin is not released. Similarly, some of the hormones that are released from the small intestine by products of food digestion (especially fat), in particular glucagon and secretin, suppress acid secretion, although intact vagus nerves seem to be necessary for this to take place because the suppression of gastrin secretion is not observed after vagus nerves are severed (vagotomy).

The intestinal phase

The gastric chyme that enters the duodenum and jejunum continues to elicit acid secretion for many hours, although the amount of acid released diminishes progressively during the digestion and absorption processes in the small intestine. Some of the actions of the intestinal phase are due to gastrin released from the duodenum, but there is evidence that another hormonelike substance not yet characterized may be responsible. Finally, as certain amino acids and small peptides are infused into the circulation during this phase, they promote gastric acid secretion. It is possible, therefore, that the absorption of the products of protein digestion also may have a role in the intestinal phase. The phases of gastric secretion overlap, and there is an interrelation and some interdependence between the neural and humoral pathways.

Gastric digestion. The stomach is not primarily a digestive organ and is not essential to life. It is possible, for example, after total gastrectomy (the complete removal of the stomach) for a person to remain, or to become, obese because most of the digestion and absorption of food takes place in the intestines. The main functions of the stomach are to commence the digestion of carbohydrates and proteins, to convert the meal into chyme, and to discharge the chyme into the small intestine periodically as the physical and chemical condition of the mixture is rendered suitable for the next phase of digestion.

Salivary amylase works on food starch while the acidity of the mixture is low, around pH6, but ceases when the acidity of the mixture increases with greater acid secretion. The gastric pepsins account for only about 10 to 15 percent of the digestion of protein and are most active in the first hour of digestion. A lipase (fat-splitting enzyme) may be present in gastric juice, but it is not capable of digesting medium- and long-chain fatty acids, and the proportion of short-chain fatty acids in the food is small. Thus, little fat digestion proceeds in the stomach.

In addition to pepsin, the gastric juice of some mammals (*e.g.*, calves) contains another enzyme, rennin, which clumps milk proteins, thus taking them out of solution and making them more susceptible to the action of a proteolytic enzyme. Few animals other than mammals have such an enzyme. Although rennin aids digestion, it is not itself a digestive enzyme since it does not catalyze a hydrolytic reaction.

Gastric absorption. Although the stomach absorbs few of the products of digestion, it can absorb many other substances, including glucose and other simple sugars, amino acids, and some fat-soluble substances. Water moves freely from the gastric contents across the gastric mucosa into the blood. The net absorption of water from the stomach is small, however, because water moves just as easily from the blood across the gastric mucosa to the lumen of the stomach. About 60 percent of the "heavy" water (isotopic) that is placed in the stomach to trace its movements is absorbed into the blood in 30 minutes.

Alcohol and drugs

A number of alcohols, including ethyl alcohol, are readily absorbed from the stomach. The membranes of the cells that line the stomach are partly composed of lipids; hence items in the diet that are fat-soluble may penetrate them. The pH of the gastric contents also determines whether or not some substances are absorbed. At a low pH, for example, the environment is acidic and aspirin is absorbed from the stomach almost as rapidly as water, but as the pH of the stomach rises and the environment becomes more basic, aspirin is absorbed more slowly.

Nevertheless, the absorption of water and alcohol can be slowed if the stomach contains foodstuffs and especially fats, probably because gastric emptying is delayed by fats, and most water in any situation is absorbed from the jejunum. The absorption of drugs is influenced by the rate of gastric emptying, the presence of food or other drugs, the particular formulation of the drug, and the vehicle carrying it. The rate of emptying of the stomach depends upon the physical and chemical composition of the meal. Fluids empty more rapidly than solids, carbohydrates more rapidly than proteins, and proteins more rapidly than fats. Liquid meals that have the same osmotic pressure as blood (isotonic) empty more rapidly than hypotonic or hypertonic mixtures.

Small intestine. The small intestine is the principal digestive organ of the human alimentary canal. Three principal activities within the small intestine are especially adapted for its role in digestion: (1) its motor activity allows intraluminal contents to be mixed and transported; (2) its secretions into the lumen provide the necessary enzymes and other constituents essential for normal digestion; and (3) its absorptive capabilities are highly selective. The absorptive functions of the small intestine are made possible by the specialized structures of the intestinal mucosa that line the small intestine.

Motor activity. The contractions of the spiral muscle layers of the small intestine (the circular and the longitudinal muscles) are regulated by an electrical impulse that begins with the passage of calcium ions into the muscle cell. The depolarization of the muscle cell membranes, or an excess of positive charges on the inside of the cell, cause the myofibrils (the contracting components of the myofilaments that constitute the muscle tissues) to contract. The two spiral layers of muscle then contract, causing the motor activity that permits the mixing and transporting of the food in the small intestine. The rate of these contractions is governed by the rate of depolarization of the muscle cell membrane.

The part played by the hormones motilin and neurotensin (see below *Hormonal control of gastrointestinal function*) remains uncertain, although when they are infused intravenously in experimental studies, they promote

motor activity in both the small and large intestines. The basic neural control of the parasympathetic nervous system, which induces smooth muscle contraction, is exerted by the cholinergic neurons that secrete acetylcholine. The basic neuronal control of the sympathetic system, which induces relaxation of the smooth muscle wall, is exerted by the adrenergic neurons, which secrete epinephrine. In the myenteric plexus (a network of nerve fibres in the wall of the intestine) there are several other messenger substances capable of modulating smooth muscle activity, including somatostatin, serotonin (5-hydroxytryptamine), and the enkephalins. With at least seven such substances in and around the smooth muscle, there is some confusion as to their respective roles.

Types of motor activity. The primary purposes of the movements of the small intestine are to provide mixing and transport of intraluminal contents. Two types of motor activity have been recognized: segmenting contractions and peristaltic contractions. A characteristic of small intestine motility is the inherent ability of the smooth muscle constituting the wall of the intestine to contract spontaneously and rhythmically. This phenomenon is independent of any extrinsic nerve supply to the small intestine and creates pressure gradients from one adjacent segment of the intestine to another. The pressure gradients, in turn, are primarily responsible for transport within the small intestine.

Segmentation

The predominant motor action of the small intestine is the segmenting contraction, which is a localized circumferential contraction, principally of the circular muscle of the intestinal wall. Segmenting contractions mix, separate, and churn the intestinal chyme. The contraction involves only a short segment of the intestinal wall, less than one to two centimetres, and constricts the lumen, tending to divide its contents. There is a gradual decrease in the number of segmenting contractions as the chyme moves from the duodenum to the ileum. This has been described as the "gradient" of small intestine motility. Although segmenting contractions usually occur in an irregular manner, they can occur in a regular or rhythmic pattern and at a maximum rate for that particular site of the small intestine (rhythmic segmentation).

Rhythmic segmentation may occur only in a localized segment of small intestine, or it may occur in a progressive manner, with each subsequent segmenting contraction occurring slightly below the preceding one (progressive segmentation).

The duodenal pacemaker sends electrical impulses down the small intestine at a rate of 11 cycles per minute in the duodenum, gradually decreasing to eight cycles per minute in the ileum. These electrical changes are propagated down the small intestine in the longitudinal muscle layer of the wall of the small intestine. Superimposed upon the slow-wave electrical activity may be fast, spikelike electrical changes. This type of electrical activity originates in the circular muscle layer of the intestinal wall and occurs when the circular layer contracts to form a segmenting contraction. Between meals electrical activity is random, but it becomes intense when at 80- to 120-minute intervals a major migrating complex rolls along at six to eight centimetres per second.

Peristaltic contractions

A peristaltic contraction may be defined as an advancing ring, or wave, of contraction that passes along a segment of the gastrointestinal tract. It normally occurs only over a short segment (approximately every six centimetres) and moves at a rate of about one or two centimetres per minute. This type of motor activity in the small intestine is used primarily for transport of intraluminal contents downward, usually one segment at a time.

When some inflammatory condition of the small bowel exists, or when irritating substances are present in the intraluminal contents, a peristaltic contraction may travel over a considerable distance of the small intestine: this has been termed the peristaltic rush. Diarrhea due to common infections is frequently associated with peristaltic rushes. Most cathartics produce their diarrheic effect by irritating the intestinal mucosa or by increasing the contents, particularly with fluid.

Digestive secretions. There are many sources of digestive secretions into the small intestine. The gastric chyme that is emptied into the duodenum contains gastric secretions that will continue their digestive processes for a short time in the small intestine. One of the major sources of digestive secretion is the pancreas, a large gland that produces both digestive enzymes and hormones. The pancreas empties its secretions into the duodenum through the major pancreatic duct in the papilla of Vater and the accessory pancreatic duct a few centimetres away from it. Pancreatic juice contains enzymes that digest proteins, fats, and carbohydrates. Secretions of the liver are delivered to the duodenum by the common bile duct, via the gallbladder, and are received through the papilla of Vater. Secretion from the mucosal surface of the small intestine is minimal and usually is no more than a few millilitres per hour for any one segment of intestine.

The composition of the fluids secreted (the succus entericus) varies somewhat in different parts of the intestine. In the duodenum, for example, where the mucous Brunner's glands are located, the secretion contains more mucus. In general, the secretion of the small intestine is a thin, colourless or slightly straw-coloured fluid, containing flecks of mucus, water, inorganic salts, and organic material. The inorganic salts are those commonly present in body fluids, with the bicarbonate concentration higher than it is in blood. Aside from mucus, the organic matter consists of enzymes and cellular debris. Many enzymes have been reported to be present in intestinal secretions. These include a pepsin-like protease (from the duodenum only), an amylase, a lipase, at least two peptidases, sucrase, maltase, enterokinase, alkaline phosphatase, nucleophosphatases, and nucleocytases.

Composition of intestinal secretions

Secretions into the small intestine are controlled by nerves (vagus) and hormones. Except in the duodenum, the quantity of the fluid secreted is never great, even under conditions of stimulation. Stimulation of the vagus nerves in experimental animals causes a moderate secretion of juice from the duodenum and smaller amounts from the rest of the small intestine after a latent period of one to one and one-half hours. Stimulation of sympathetic nerves supplying the small intestine causes no secretion, but cutting the vagus nerves results in an increase in secretion. The most effective stimuli for the secretion of succus entericus are local mechanical or chemical stimulations of the intestinal mucous membrane. Such stimuli always are present in the intestine in the form of chyme and food particles. A hormone, enterocrinin, extracted from intestinal mucosa in humans, causes the secretion of juice from the intestinal mucosa in animals.

The three principal components in the human diet that require digestion are carbohydrates, proteins, and fats. Most of the digestive processes that solubilize these substances and reduce them to relatively simple organic compounds occur in the upper (proximal) small intestine. The plasma membrane over the microvilli is thicker and richer in protein and lipids than is the plasma membrane on the cells at the side and base of the villus. Fused to this brush border is the layer of glycoprotein that is known as the "fuzzy coat." The microvilli are important to both digestion and absorption. They elaborate the enzymes, disaccharidases and peptidase, that hydrolyze disaccharides and polypeptides to monosaccharides and to dipeptides and amino acids, respectively.

Microvilli

Molecular receptors for specific substances are found on the microvilli surfaces at different levels in the small intestine. This may account for the selective absorption of particular substances at particular sites, for example, intrinsic-factor-bound vitamin B_{12} in the terminal ileum. Such receptors may also explain the selective absorption of iron and calcium in the duodenum and upper jejunum. Furthermore, there are transport proteins in the microvillus membrane associated with the passage of sodium ions, D-glucose, and amino acids.

Actin is found in the core of the microvillus, and myosin is found in the brush border; because contractility is a function of these proteins, the microvilli have motor activity that presumably initiates the stirring and mixing actions within the lumen of the small intestine.

Certain nutrients are partly digested in the glycoprotein

fuzz. The products of this digestion penetrate the surface of the microvillus and pass into the interior of the enterocytes (intestinal epithelial cells) for further digestion, and ultimately they are transferred to the intercellular space.

Water and solutes pass through pores in the surface epithelium of the mucosa. The size of the pores is different in the ileum than in the jejunum; this difference accounts for the various rates of absorption of water at the two sites. The enterocytes are joined near their apex by a contact zone known as a "tight junction." These junctions are believed to have the pores that are closed in the resting state and dilated when absorption is required.

The "tight junction"

The movement of sodium and chloride is dependent to an extent on solvent drag; *i.e.*, solutes are caught up in a moving stream of water and the water movement causes an increased concentration of the solute on the side of the membrane from which the water originally had come. This gradient dictates the direction of solute movement. This is one mechanism for solute absorption, the other being active transport systems that move solutes across the cell membrane by the expenditure of energy.

Large intestine. Unlike that of the small intestine, the mucosa of the large intestine (colon) does not have villi. The mucosa of the large intestine is more glandular. Its surface is covered by tall columnar epithelial cells, and it has many crypts that are lined with mucous glands. The principal substances secreted by the colon are mucus, potassium, and bicarbonate. The mucus aids in lubricating the intestinal contents and facilitates their transport during mass movement contractions.

The primary function of the colon is to absorb water, sodium, and chloride from the chyme. Each day approximately 1.5 to two litres of fluid chyme pass through the ileocecal valve that separates the small and large intestines. The fluid is reduced by absorption in the colon to around 150 millilitres. The residual indigestible matter, together with sloughed-off mucosal cells, dead bacteria, and food residues not digested by bacteria, constitute the feces.

The rate of transit depends largely on the frequency and strength of the contractions of the smooth muscle in the colon wall. The inner layer of muscle is wound in a tight spiral around the colon, and contraction results in segmentation of the lumen and its contents: the spiral of the outer layer follows a loose undulating course, and contraction causes the contents of the colon to shift forward and backward. The bulk of the contents influences these muscular activities, in particular the amount of undigested fibrous material (roughage).

Other colonic functions include the excretion of surplus bicarbonate and potassium and the absorption of sodium and chloride. Thus, the colon has a role in maintaining the osmolality, or level of solutes, of the blood, which is partly dependent on the concentration of electrolytes. In vertebrates, these various functions are controlled by neural and hormonal mechanisms. The colon contains large numbers of bacteria, and a state of symbiosis exists (*i.e.*, disparate groups living together to mutual advantage). The bacteria synthesize niacin (nicotinic acid) and thiamine (vitamin B_1), two vitamins that are essential to several metabolic activities as well as to the function of the central nervous system. The bacteria also synthesize vitamin K, which is essential to the formation of the factors that are responsible for the clotting ability of the blood.

Symbiosis

Colonic motility. The electrical activity of the muscles of the colon is more complex than that of the small intestine. In the lower (distal) half of the colon, variations from the basic rhythmic movements are present 5 to 20 percent of the time, and in the rectum, variations are present 70 to 90 percent of the time. Slow-wave activity that produces contractions from the ascending colon to the descending colon occurs at the rate of 11 cycles per minute, and slow-wave activity in the sigmoid colon and rectum occurs at a rate of six cycles per minute. Electrical spike potentials, which indicate local contractions, migrate distally in the colon at the rate of four centimetres per second. Retrograde, or reverse, movements occur mainly in the upper (proximal) colon. The local contractions and the retrograde propulsions ensure mixing of the contents and good contact with the mucosa.

Neuronal and hormonal mechanisms combine to elicit colonic motility, but the proportionate contribution of each mechanism remains uncertain. Whereas the presence of fat in the colon stimulates motility, the presence of carbohydrate and protein do not. Unabsorbed bile salts and bile acids also increase motor activity. The peptide hormones gastrin and cholecystokinin stimulate colonic motility; secretin, glucagon, and vasoactive intestinal peptide suppress it. Mastication stimulates colonic motor activity.

The accumulating feces are stored in the descending and sigmoid sectors. Once or twice in 24 hours a mass peristaltic movement shifts the feces onward. The rectum is normally empty, but when it is filled with gas, liquids, or solids to the extent that the intraluminal pressure is raised to 15 to 20 centimetres of water, the impulse to defecate occurs.

The act of defecation is preceded by a voluntary effort, which, in turn, probably gives rise to stimuli that augment the visceral reflexes, although these originate primarily in the distension of the rectum. Centres that control defecation reflexes are found in the hypothalamus of the brain, in two regions of the spinal cord, and in the ganglionic plexus of the intestine. As the result of these reflexes, the internal anal sphincter relaxes.

Defecation reflexes

The mass contraction of the colon carries its contents into the pelvic colon, which in turn transfers them into the rectum, eventually to be evacuated by way of the anus. Thus, the entire distal colon from the splenic flexure to the anus may be emptied at one time. A prominent mechanical feature of defecation is the contraction of the longitudinal muscles of the distal and pelvic colon. The resulting shortening of the distal colon tends to elevate the pelvic colon and obliterates the angle that it normally makes with the rectum. The straightening and shortening of the passage facilitates evacuation.

Anatomy of defecation. The anal canal between the rectum and the body surface is three to four centimetres long. It is surrounded by the muscles constituting the internal and external sphincters. The inner ring of muscle is smooth in type and controlled by the autonomic system. The outer ring is a complex formed from the puborectal fibres of the levator and muscles and an external sphincter; both of these are striated muscle and can be relaxed or contracted voluntarily.

The musculus puborectalis forms a sling around the junction of the rectum with the anal canal, and it is maintained in a constant state of tension (tone). This results in an angulation of the lower rectum so that the lumen of the rectum and the lumen of the anal canal are not in continuity, a feature essential to continence. Continuity is restored between the lumina of the two sectors when the sling of muscle relaxes, and the angulation is overcome.

The anal canal is lined proximally with columnar mucosa innervated by the autonomic system and distally by squamous cells (a modification of skin) that are innervated by peripheral nerves.

Fibre. Fibre includes cellulose, hemicellulose, pectins, gums, and lignins, mostly polysaccharides built from five-carbon (furanose) or six-carbon (pyranose) rings. Fibres are polymers (compounds of high molecular weight formed by the linear combination of simpler repeating molecules, or monomers) with various linkage arrangements between the molecules. They defy the hydrolyzing enzymes of the small intestine. In the cecum and colon, however, the commensal bacteria break down the linkages.

Fibre can hold water either by binding to its surface or by trapping it in interstices in a spongelike manner. Fibre also has a capability for ion exchange and adsorption. Ion exchange is promoted by the electrical potential at the interface between the water and the fibre, which thus gives it the property of a colloid. As the bacterial enzymes break down the polysaccharides these cause disintegration of the matrix. Adsorption can take place on the surface of fibre as well. Drugs prescribed for the treatment of a disease may be adsorbed onto the fibre and become unavailable to the body. The adsorption of bile acids and other digestive products by fibre may not necessarily be disadvantageous.

(N.C.H./W.S.)

The hepatocyte

Liver. Hepatocytes occupy about 80 percent of the volume of the liver, and their cytoplasm (the area surrounding the nucleus) contains many mitochondria, which provide the energy needed for the many synthetic and metabolic functions of the liver cell. The cytoplasm also contains a series of long tubules, called the endoplasmic reticulum, which contains many enzymes essential to liver function. Some of the membranes of the endoplasmic reticulum appear granular, or rough, owing to the presence of ribosomes, which are responsible for forming specific polypeptide (protein) chains from single amino acids. The non-ribosomal, or smooth, endoplasmic reticulum is the site where cytochromes (combinations of heme from hemoglobin with various proteins) and certain enzymes undertake the important hepatic functions of drug and hormonal metabolism, cholesterol synthesis, and conjugation with carbohydrate moieties of bilirubin and other fat-soluble metabolic and foreign compounds, thereby making them soluble in water. The Golgi apparatus, a series of tubular structures between the endoplasmic reticulum and the canaliculus, acts as a transport station for newly made proteins and other hepatocytic products before they are conveyed to other parts of the cell or out of the cell entirely. Lysosomes, another important cytoplasmic constituent, are responsible for the intracellular storage of pigments, such as iron or copper, and for the digestion of certain contents, such as glycogen or foreign particles. The nucleus of the hepatocyte has no apparent functions that distinguish it from nuclei of other cells. The nucleus guides replication of the cell and transmits genetic material in the form of messenger ribonucleic acid (mRNA) from deoxyribonucleic acid (DNA) to organelles located in the cytoplasm.

Major functions of the liver

The major functions of the liver are to participate in the metabolism of protein, carbohydrate, and fat; to synthesize cholesterol and bile acids; to initiate the formation of bile; to engage in the transport of bilirubin; and to metabolize and transport certain drugs. Dietary amino acids from ingested protein are circulated to the hepatocytes of the liver, where they are either cycled into the production of specific human proteins, or used as a source of energy after having had the amino group removed (deamination) and converted into glucose through a process called gluconeogenesis. The ammonia released from amino acids in the process of gluconeogenesis is converted into urea only in the hepatocyte by way of a special enzyme-controlled metabolic process called the urea cycle. Except for the immunoglobulins (gamma globulins), which are produced in the spleen and lymphoid tissue, hepatocytes are the major source for the production of all blood proteins. These include albumin, proteins essential to the coagulation of blood, certain enzyme inhibitors such as alpha$_1$-antitrypsin, lipoproteins, and the transport proteins such as thyroxine-binding globulin, ceruloplasmin (copper), transferrin (iron), and transcobalamin (vitamin B_{12}). The stability of these proteins is quite variable; albumin, for example, has a half-life of more than 15 days, while some of the blood coagulation factors are functional for only seven hours.

The liver controls the transport and storage of energy-producing carbohydrates. Glucose, which is one of the two monosaccharide components of table sugar (sucrose) and milk sugar (lactose) and is the sole building block of dietary polysaccharides, such as starch, is combined with phosphate in the liver cell and either transported to peripheral tissues for metabolic purposes or stored in the hepatocyte as glycogen, a complex polysaccharide. Specific enzyme systems are present in the hepatocyte for these conversions, as well as for the translation of other dietary monosaccharides (fructose from sucrose and galactose from lactose) into glucose. The hepatocyte is also able to convert certain amino acids and products of glucose metabolism (pyruvate and lactate) into glucose through gluconeogenesis.

The liver also plays a central role in metabolizing fat by converting stored fatty acids to their energy-releasing form, acetylcoenzyme A (acetyl CoA), when hepatic glucose and glycogen stores are exhausted or unavailable for metabolic purposes (as in diabetic ketoacidosis). The liver also plays a role in the formation of storage fats (triglycerides) whenever dietary carbohydrates, protein, or fat exceeds the requirements of tissues for glucose or the needs of the liver for glycogen. The liver also synthesizes cell membrane components (phospholipids) and proteins (lipoproteins) that carry lipids (fats and cholesterol) in the blood.

Cholesterol

Cholesterol is a major factor in cell membrane structure, and it is essential to cellular survival. Cholesterol is a four-ringed sterol that is absorbed from the diet or synthesized from dietary acetyl CoA by the liver and the intestinal lining. Excess cholesterol is a major constituent of the bile that is produced in the liver and transmitted into the intestine. About half of the hepatic cholesterol is first converted into bile acids, which have the same sterol structure as cholesterol but contain three fewer carbon atoms and an acidic side chain. The rate of bile flow is controlled in part by the rate of bile acid secretion by the hepatocyte.

The hepatocyte also acts to transport certain water-insoluble products of metabolism and agents foreign to the body into the bile and urine. Bilirubin is the product of hemoglobin metabolism after the iron and protein fractions have been removed. Bilirubin is formed in the bone marrow and the lymphatic tissue and is carried to the liver after becoming bound to plasma albumin. It is released at the hepatocytic sinusoidal membrane and is transported to the smooth endoplasmic reticulum where, in an enzymatic system, it is conjugated with one or two molecules of glucuronic acid, thereby becoming soluble in water and excretable in bile. Similarly, many drugs that are normally insoluble in water are conjugated, or joined with other substances to detoxify them, in phase II reactions in hepatocytes after having been oxidized by cytoplasmic enzyme systems (phase I reactions). Small drug metabolites are excreted into the urine, while larger molecules leave the body through the bile and the feces.

Biliary tract. Aside from inorganic ions (sodium, potassium, calcium, magnesium, chloride, and bicarbonate), bile contains protein and bilirubin; the latter is responsible for its golden colour in dilute solutions and dark amber colour in concentrate. It is richest, however, in bile acids (derived from cholesterol in the hepatocyte), phospholipids (largely phosphatidyl choline, or lecithin), and cholesterol. Normally the cholesterol, which is not soluble in watery secretions, is carried in a colloidal solution in bile in the form of mixed aggregates of complexes containing bile acids and lecithin. In the absence of adequate amounts of lecithin and bile acids, cholesterol crystallizes.

Types of bile acid

The liver synthesizes two types of primary bile acid from cholesterol, called chenodeoxycholic acid and cholic acid. In the lower intestine, bacterial action removes one of the hydroxyl groups (dehydroxylation) from cholic acid, changing it to deoxycholic acid. This secondary bile acid appears in bile because it is absorbed from the intestine and recirculated to the liver. Chenodeoxycholic acid is also dehydroxylated in the intestine, becoming lithocholic acid, a small amount of which is also reabsorbed and appears in normal bile. Neither deoxycholate nor lithocholate appears to be an important factor in the incorporation of cholesterol micelles, or vesicles, in bile because they lack the important feature of carrying hydroxyl groups at both the C_3 and C_7 positions of the sterol nucleus.

The total bile acid pool at any one time measures about three grams, almost all of which is contained at rest in the gallbladder. It is maintained largely because about 95 percent of the bile acids entering the intestine from the biliary system are reabsorbed actively in the lower portion of the small intestine, so that only 0.2 to 0.6 gram of bile acids is lost daily. This loss can be replaced readily by the normal liver.

Bile is formed initially in the hepatocyte, and the rate of formation is dependent primarily on the rate at which bile acids are secreted into bile canaliculi. A portion of the flow of bile, however, is related to factors other than the secretion of bile acids. This flow appears to be dependent on the secretion of sodium from the hepatocyte and is also partially governed by the action of intestinal hormones such as secretin, cholecystokinin, and gastrin. In its passage through the biliary tract, hepatic bile is con-

centrated to as little as one-tenth of its original volume by the selective reabsorption of water, chloride, and bicarbonate. This concentration process takes place largely in the gallbladder, with the result that bile from this organ is much thicker in density and darker in colour (due to concentration of pigments) than is bile emerging from the liver. Distension of the duodenum, particularly by a meal containing fat, provokes the secretion of CCK, a hormone that causes contractions of the muscular layer in the wall of the gallbladder. Bile concentration is greatly reduced if the gallbladder is removed, but this effect apparently does not impede the primary digestive function of bile, which is to aid in the dispersion and digestion of fat in the lumen of the intestine.

Acinar cells

Pancreas. The acinar cells constitute more than 95 percent of the cellular population of the exocrine pancreas. Their membranes contain specific receptor sites for cholinergic (parasympathetic) and polypeptide (secretin and cholecystokinin) agonists. Acetylcholine and cholecystokinin (CCK) bind to parasympathetic receptor sites and, in the presence of calcium, cause marked enzymic secretion by acinar cells. Binding of vasoactive intestinal polypeptide (VIP) or secretin to acinar receptors activates the adenosine monophosphate shunt and causes increased production of bicarbonate, sodium, water, and enzymes by acinar cells and small ductal cells. Bicarbonate is secreted in exchange for chloride, and sodium is exchanged for hydrogen, with a resultant increased acidity of the blood leaving the actively secreting pancreas.

In the acinar cell, almost all enzymatic proteins are synthesized on ribosomes from amino acids carried to the pancreas by the bloodstream. The enzyme precursors are conjugated in the Golgi apparatus, and then concentrated into membrane-wrapped cytoplasmic zymogen granules. Upon stimulation with CCK or acetylcholine, these zymogen granules migrate to the apex of the acinar cell where they are extruded into the central ductal lumen. These stimulants also increase the synthesis of zymogen granules. In the absence of CCK and acetylcholine, as in fasting subjects or in patients being fed intravenously, the synthesis of zymogen by the acinar cells is markedly reduced. Pancreatic atrophy also occurs after removal of the pituitary gland, probably due to the absence of growth hormone. Thus CCK, acetylcholine, and growth hormone are pancreatotrophic hormones. The pancreas itself also appears to secrete an as yet unidentified hormone that is trophic, or nutritive, to the liver.

Pancreatic enzymes

Acinar cells produce a great variety of digestive proteins, or enzymes, involved principally with the degradation in the intestine of dietary proteins (proteases), fats (lipases), and carbohydrates (amylases). Other protein secretions include a trypsin inhibitor, a so-called stone protein that keeps calcium in solution, and various serum proteins, including albumin and immunoglobulins. Proteases are of two types, endopeptidases and exopeptidases. The former, including trypsinogen, chymotrypsinogen, and proelastase, split food proteins at internal amino acid linkages, while the latter, including procarboxypeptidases A and B and kallikreinogen, cleave terminal chain residues from polypeptides (proteins). Each of these proteases is secreted in an inactive, or precursor, form. Trypsinogen is activated in the intestine by enterokinase, an enzyme liberated from duodenal lining cells by the interaction of bile acids and CCK. This activation of trypsinogen to trypsin is brought about by the cleavage from it of six terminal amino acid residues. The other proteases are activated by free trypsin. The net effect of these proteases is to reduce dietary proteins to small polypeptides (two to six amino acid chains) and to single amino acids. Lipases include phospholipase, esterase, colipase, and lipase, which function to reduce neutral fats (triglycerides) and free fatty acids and monoglycerides, to hydrolyze dietary phospholipids, and to reduce long-chain fatty acids. Lipases require the presence of bile acids in the intestinal lumen for the formation of micellar solutions of fat prior to optimal digestion. There are at least six varieties of amylase in pancreatic juice whose function is to hydrolyze complex carbohydrates (polysaccharides) to disaccharides and trisaccharides.

Pancreatic secretion is mediated by stimulants such as secretin, a hormone released from the duodenal wall by the physiological introduction of gastric acid into the duodenum, and CCK, which is released by the presence of dietary fat and amino acids, as well as by gastric hydrochloric acid. Stimulation of the vagus nerve by hunger also releases acetylcholine in the acinar milieu and leads to enzymatic secretion. The flow induced by secretin is a high-volume, bicarbonate-and-water mixture that also contains moderate amounts of enzymes. The flow induced by CCK is thick, low in bicarbonate and highly concentrated in enzymes. (H.J.Dw.)

Features of the gastrointestinal tract

GENERAL FEATURES OF DIGESTION AND ABSORPTION

Methods of absorption

There are four means by which digestive products are absorbed: active transport, passive diffusion, facilitated diffusion, and endocytosis.

Active transport involves the movement of a substance across the membrane of the absorbing cell against an electrical or chemical gradient. It is carrier-mediated, that is, the substance is temporarily bound to another substance that transports it across the cell membrane, where it is released. The process requires energy and is at risk of competitive inhibition by other substances; that is, other substances with a similar molecular structure can compete for the binding site on the carrier. Passive diffusion requires neither energy nor carrier; the substance merely passes along a simple concentration gradient from an area of high concentration of the substance to an area of low concentration, until a state of equilibrium exists on either side of the membrane. Facilitated diffusion also requires no energy, but it involves a carrier, or protein molecule located on the outside of the cell membrane that binds the substance and carries it into the cell. The carrier may be competitively inhibited. Endocytosis takes place when the material to be absorbed, on reaching the cell membrane, is enfolded into it. That part is then pinched off into the cell interior. This process is similar to phagocytosis.

Absorption of all food by the small intestine occurs principally in the jejunum; however, the duodenum, although the shortest portion of the small intestine, has an extremely important role. The duodenum receives not only chyme saturated with gastric acid, but pancreatic and liver secretions as well. It is in the duodenum that the intestinal contents are rendered isotonic with the blood plasma; *i.e.*, the pressures and volumes of the intestinal contents are the same as those of the blood plasma, so that the cells on either side of the barrier will neither gain nor lose water.

The bicarbonate secreted by the pancreas neutralizes the acid secreted by the stomach. This brings the intestinal contents to the optimal pH, allowing the various enzymes to act on their substrates at peak efficiency. A number of important gastrointestinal hormones regulate gastric emptying, gastric secretion, pancreatic secretion, and contraction of the gallbladder. These hormones, along with neural impulses from the autonomic nervous system, provide for autoregulatory mechanisms for normal digestive processes.

Electrolytes and water

Most salts and minerals, as well as water, are readily absorbed from all portions of the small intestine. Twelve to 25 grams of sodium are present in the average daily diet. The sodium is absorbed by an active process, and the necessary metabolic energy is provided by the epithelial cells of the mucosa of the small intestine. Sodium is moved from the lumen of the intestine across the mucosa against a concentration gradient (*i.e.*, a progressive increase in the concentration of sodium) and an electrochemical gradient (*i.e.*, a gradual increase in the concentration of charged ions). Sodium ions are absorbed more readily from the jejunum than from other parts of the small intestine.

Potassium is absorbed at about 5 percent of the rate of sodium. It is thought that potassium moves across the intestinal mucosa passively or by facilitated diffusion as a consequence of water absorption. Chloride is readily absorbed in the small intestine and probably takes place as a consequence of sodium absorption. The absorption of water appears to be secondary to the absorption of electrolytes (substances that dissociate into ions in a solution). Water absorption occurs throughout the small intestine,

though chiefly in the jejunum. The upper small intestine absorbs approximately 95 percent of a 50-gram sample within 10 minutes. Water moves freely across the intestinal mucosa both ways, but tends to move in the direction of the hypertonic solution (the solution into which a net flow of water occurs), and away from the hypotonic solution (one from which a net flow of water occurs). Thus, if the contents of the lumen are hypotonic, water moves rapidly from the lumen to the blood. If the contents of the intestinal lumen are hypertonic, water moves more rapidly from the blood into the lumen. This two-way movement of water tends to maintain the intestinal contents in an isotonic state.

SPECIFIC FEATURES OF DIGESTION AND ABSORPTION

Carbohydrates. Carbohydrates are absorbed as monosaccharides (simple sugars that cannot be further broken down by hydrolysis) or disaccharides (carbohydrates that can be hydrolyzed to two monosaccharides). The absorption of glucose and galactose is dependent on the presence of sodium and uses active transport to move against a concentration gradient. Amylose, a starch polysaccharide (a carbohydrate that contains many monosaccharides), accounts for 20 percent of dietary carbohydrate. It consists of a straight chain of glucose (sugar) molecules bound to their neighbours by oxygen links. The bulk of the starch is amylopectin, which has a branch chain linked in after every 25 molecules of glucose on the main chain.

Digestion of starch

Only a small amount of starch is digested by salivary amylase; most is rapidly digested in the duodenum by pancreatic amylase. But even this enzyme has little effect on the branch chains of amylopectin, and even less on the linkages in cellulose molecules. This accounts for the inability of humans to break down cellulose. Pancreatic amylase changes amylose to maltose (a disaccharide) and the amylopectins to dextrins.

In the brush border (comprising ultrafine microvilli) and the surface membrane of the epithelial enterocytes are the disaccharidase enzymes, lactase, maltase, sucrase, and trehalase, which hydrolyze maltose and the dextrins to the monosaccharides glucose, galactose, and fructose.

Fructose appears to be absorbed by simple diffusion, but glucose and galactose are transported by an energy-using process, probably binding it to a specific protein carrier with attached sodium ions; the sugar is released inside the enterocyte, sodium is pumped out, and the sugars then diffuse into the circulation down a concentration gradient.

Proteins. The digestion of protein entails breaking the complex molecule first into peptides, each having a number of amino acids, and second, into individual amino acids. The pepsins are enzymes secreted by the stomach in the presence of acid that breaks down proteins (proteolysis). The trypsins (proteolytic enzymes secreted by the pancreas) are much more powerful than pepsins so that the greater part of protein digestion occurs in the duodenum and upper jejunum. Hence, even after total removal of the stomach, protein digestion usually is not impaired.

Pancreatic secretion contains inactive protease precursors that become enzymatically active after being mixed with another enzyme, enterokinase, which is secreted from the microvillous component of the enterocytes in the duodenal and jejunal mucosa. Enterokinase converts an inactive precursor, trypsinogen, to trypsin, which activates the other pancreatic proteases, including chymotrypsin and elastase. Trypsin, chymotrypsin, and elastase are known as endopeptidases and are responsible for the initial breakdown of the protein chains to peptides by hydrolysis. The next step, the breakdown of these peptides to smaller molecules and then to individual amino acids is brought about by the enzymic activity of carboxypeptidases, which are also secreted by the pancreas.

Endopeptidases

Peptidase activity commences outside the enterocytes (in the mucus and brush border) and continues inside the cell. There appears to be a different peptidase for each stage of the breakdown of protein to amino acids. Likewise, the transport of different peptides involves different mechanisms. Dipeptides (peptides that release two amino acids on hydrolysis) and tripeptides (peptides that release three amino acids) are moved from the surface brush border into the cell by an energy-requiring process involving a carrier protein. Small peptides with few amino acids are absorbed directly as such. The greater part of the breakdown of peptides to amino acids takes place within the enterocyte. Curiously, small peptides are absorbed more rapidly than amino acids, and indeed the precise details of the mechanism for absorption of amino acids is largely unknown. It is known that some amino acids have a specific individual transport system while others share one.

Transport of peptides

Amino acids may be classified into groups, depending upon their optical rotatory characteristics (*i.e.,* whether they rotate polarized light to the left, or levo; or to the right, or dextro) and in terms of reactivity, or acidity (pH). Levorotatory amino acids are absorbed extremely rapidly—much more rapidly than are dextrorotatory amino acids. In fact, levorotatory amino acids are absorbed almost as quickly as they are released from protein or peptide. Neutral amino acids have certain structural requirements for active transport, and if these specific structural arrangements are disturbed, active transport will not occur. Basic amino acids, which have a pH above 7, are transported at about 5 to 10 percent of the rate of neutral levorotatory amino acids.

Fats. Almost all dietary fat is stored as triglycerides. Solubility in water is necessary in order for fat to be transferred from the lumen of the intestine to the absorptive cells. Many factors, such as the length of the fatty acid chains of the triglycerides, play an important role in determining this solubility. Triglycerides have three long chains of fatty acids (LCFA) attached to a glycerol framework, and they are insoluble in water. The remainder are medium-chain triglycerides (MCT), which can be absorbed intact by the mucosa of the small intestine. Lipase acts on MCT's to free medium-chain fatty acids (MCFA), which are more soluble in water than the LCFA's and move quickly through the cells and pass into the portal circulation and then to the liver. Because MCFA's are more soluble in water, and because MCT's are absorbed by the mucosa, the addition of MCT's to the diet when disease or surgery impair enzyme hydrolysis of fat guarantees that some fat will be available for nutrition.

Pancreatic lipase

Long-chain fatty acids attached to the triglycerides are attacked by the pancreatic enzyme lipase. Two of the three fatty acid chains are split off, leaving one attached to the glycerol (forming a monoglyceride). In the presence of excess levels of bile salts, however, the lipolytic activity of pancreatic lipase is inhibited. Another pancreatic enzyme, colipase, binds to the bile salts, leaving lipase available to attack the triglycerides. The monoglycerides that result from these splitting processes combine into a complex called a micelle. The micelle, because of its physiochemical properties, permits fat components to be soluble in water. Because bile salts have a hydrophobic, or water-repelling region, and a hydrophilic, or water-attracting region, the micelle is formed with bile salts arranged around the outside with hydrophobic ends facing inside, and hydrophobic fatty acids, monoglycerides, phospholipids, and cholesterol, as well as the fat-soluble vitamins A, D, E, and K, in the centre.

There is a layer of fluid overlying the surface cells of the mucosa of the small intestine known as the "unstirred" layer. It is across this layer that the micelles must pass to reach the cell membranes. The rate of diffusion through the unstirred layer is determined by the thickness of the layer and the gradient in concentrations of the various elements of the transport system from the lumen of the intestine to the cell membrane. Underneath the unstirred layer is another layer known as the fuzz, which mainly comprises mucus. Beneath the fuzz is the brush border on the surface of the cell membrane. It has a double layer of lipid which is easily penetrated by the fatty acids and monoglycerides that are soluble in lipids. Once the micelle has passed through the fuzz and the brush border, it has entered the cells of the tissues that line the intestine. The micelle disintegrates, the bile salts diffuse back into the lumen, and a carrier protein picks up the fatty acids and the monoglycerides and transports them to the endoplasmic reticulum in the cell interior. The endoplasmic reticulum is a tubular structure rich in enzymes. At this site the

Intracellular events

triglyceride is synthesized again under the influence of an enzyme catalyst called acyltransferase.

The triglycerides pass to the membrane of another tubular structure, known as the Golgi apparatus, where they are packaged into vesicles (chylomicrons). These vesicles are spheres with an outer coating of phospholipids and a small amount of apoprotein while the interior is entirely triglyceride except for a small quantity of cholesterol. The chylomicrons migrate to the cell membrane, pass through it, and are attracted into the fine branches of the lymphatic system, the lacteals, and from there pass to the thoracic duct. The whole process of absorption, from the formation of micelles to the movement out of the cells and into the lacteals, takes between 10 and 15 minutes.

The medium-chain triglycerides are broken down to medium-chain fatty acids by pancreatic lipase. Medium-chain fatty acids are soluble in water and readily enter the micelles. Ultimately, after moving across the membrane of the enterocyte, they pass into the capillary tributaries of the portal vein and then to the liver.

Fat-soluble vitamins. Fat-soluble vitamins pass with the chylomicrons into the lymphatic system. Vitamin A, first presenting as a precursor, beta-carotene, is cleaved to form retinol, which is then recombined with fatty acids before entering the chylomicron. Vitamins D and D_3 diffuse passively into the chylomicron. The absence of bile salts from the intestine, which occurs in jaundice due to obstruction of the biliary tract, severely impairs vitamin K absorption and blood clotting, with risk of hemorrhage. Vitamin E, a mixture of oils known as tocopherols, is present in hen eggs and is synthesized by such plants as soybeans, corn (maize), wheat, and palms. It passes through the enterocyte with the other lipids of the micelle and is ultimately stored in the liver. The average diet contains 15 to 25 milligrams of alpha-tocopherol per day, which covers recommended requirements.

Calcium. Calcium is required for the construction of bone; it forms part of the substance cementing together the walls of adjacent cells; and it is vital in the responsiveness to stimuli of muscle and nerve cells, which determines their excitability. The main sources of calcium are milk and milk products, meat, in which it is bound to proteins; and vegetables, in which it is bound to phytates (phytic acid) and oxalates (the salt of oxalic acid).

Sources of calcium

The absorption of calcium is influenced by conditions within the lumen of the small intestine. The acid secretion from the stomach converts the calcium to a salt, which is absorbed in the jejunum. Unabsorbed calcium is precipitated in the ileum and is excreted in the feces. Lactose, the sugar of milk, aids calcium absorption, whereas excess fatty acid and high concentrations of magnesium and oxalates interfere with it.

Calcium is absorbed across the brush border of the enterocyte cell membrane by a mechanism that requires energy. Vitamin D is essential to this process, and when deficient, the active transport of calcium stops. Other influences on absorption of calcium are the activity of the parathyroid hormone (parathormone) and of growth hormone from the pituitary gland. An average diet contains 1,200 milligrams of calcium, one-third of which is absorbed. In the passage of the blood through the kidney, 99 percent of the circulating calcium is reabsorbed. Thus, in kidney failure as well as in jejunal malabsorption states, excessive losses of calcium occur. In calcium deficiency, calcium is resorbed from the bone, thereby weakening and softening the skeletal structure.

Magnesium. An average day's diet contains around 300 milligrams of magnesium, of which two-thirds is absorbed. Half of the absorbed magnesium is excreted by the kidneys, which can regulate the amount within a range of one to 150 millimoles per day. This control is subject to the influences of the parathyroid hormone, parathormone, and the thyroid hormone calcitotonin. Magnesium is important to neuromuscular transmission. It is also an important cofactor in the enzymic processes that form the matrix of bone and in the synthesis of nucleic acid. Magnesium deficiency can result from the overuse of diuretics and from chronic renal failure, chronic alcoholism, uncontrolled diabetes mellitus, and severe intestinal malabsorption.

Magnesium has an inverse relationship with calcium. Thus, if food is deficient in magnesium, more of the calcium in the food is absorbed. If the blood level of magnesium is low, calcium is mobilized from bone. The treatment of hypocalcemia due to malabsorption includes administration of magnesium supplements.

Hematinics. Hematinics are substances that are essential to the proper formation of the components of blood. They include folic acid, vitamin B_{12}, iron, and vitamin D.

Folic acid. Folic acid (pteroylglutamic acid) is necessary to the synthesis of nucleic acids and to cell replication. In deficiency of folic acid, maturation of red blood cells (erythrocytes) is impaired. Folates are synthesized by bacteria and plants and are hydrolyzed to folic acid in the intestine. Milk and fruit are the main sources of folic acid, providing on average 500 micrograms daily, which is four times the normal requirement. The liver stores amounts up to 5,000 micrograms.

The hydrolysis of the folates, a necessary step to absorption, takes place on the brush borders of jejunal enterocytes and is completed on lysosomes (structures within the cell that contain various hydrolytic enzymes and are part of the intracellular digestive system). When hydrolysis of folates is disturbed, anemia develops. This process is interfered with by certain drugs, especially phenytoin, used in the management of epilepsy, and by the long-term use of sulfonamides in the suppression of disease. A methyl group is added to pteroylglutamic acid in the enterohepatic circulation (methylation) in the liver and is excreted in the bile. Approximately 100 micrograms are utilized each day. The method of absorption is uncertain.

Hydrolysis of folates

Vitamin B_{12}. Vitamin B_{12}, also called cobalamin because it contains cobalt, is essential to the formation of blood cells. It is a coenzyme that assists the enzymes responsible for moving folate into the cell interior. Vitamin B_{12} is a product of bacterial metabolism and enters animal tissue from this source. Although bacteria in the human colon also produce vitamin B_{12}, it cannot be absorbed at that site. Vitamin B_{12} is in a bound form in food and is liberated by proteolytic activity in the stomach and small intestine. It then binds with intrinsic factor (IF), a glycoprotein elaborated in humans by the same parietal cells that form hydrochloric acid. In some species the IF is produced in the chief cells. Intrinsic factor is essential to transport, and the B_{12} protein complex, known as transcobalamin II, is necessary to movement of the vitamin from the intestine to the rest of the body. Once the IF is attached, further proteolytic digestion of the bound vitamin is prevented. Absorption is confined to the distal 100 centimetres of ileum, especially the last 20 centimetres, where the complex binds to receptors in the brush border of the enterocytes. The process is slow; it takes three hours from its presentation in food to its appearance in the peripheral blood via the enterohepatic circulation and hepatic veins. The daily requirement of vitamin B_{12} is one microgram; however, five to 10 micrograms circulate via the ileum and liver, and the liver stores up to 5,000 micrograms. A small amount (1.5 percent) of the vitamin can be absorbed in the small intestine by diffusion.

Iron. Iron is necessary for the synthesis of hemoglobin, the oxygen-carrying compound of the red blood cells. It also has an important role as a cofactor in intracellular metabolism. The main dietary sources are meat, eggs, nuts, and seeds, and more iron is absorbed from meat (20 percent) than from vegetables (5 percent). The average daily diet contains approximately 20 milligrams of iron, and humans are unable to excrete iron that has been absorbed in excess of the daily requirement of one milligram.

Dietary sources of iron

The acid in the stomach prevents the formation of insoluble complexes, as does vitamin C. Some amino acids from dietary protein stabilize the iron in complexes of low molecular weight. Phosphates and phytates of vegetable origin, some food additives, and the inhibition of acid secretion impede the absorption of iron. Iron is almost wholly absorbed in the duodenum by a process that involves metabolic activity requiring energy. Most of the iron remains trapped in the surface enterocytes and is lost when the cells die and fall off into the intestine. The amount of iron lost seems to be related in some way to the state

of the body's iron stores, although this can be overcome if very large doses of iron are taken orally. Alcohol in the stomach and duodenum increases the rate of absorption. Transport of the iron from the enterocyte is achieved by binding to a carrier, a plasma protein called transferrin. From the intestine, it passes into the portal circulation and the liver. When the loss of iron is increased, as in excessive menstruation and in bleeding disorders, the rate of absorption is stepped up from less than one milligram per day to 1.5 milligrams or more.

Vitamin D. Vitamin D is essentially a hormone and is available from two sources. First, under the influence of photosynthesis made possible by ultraviolet rays from the Sun, a sterol compound from the liver (dehydrocholesterol) is converted to vitamin D_3. This supplies enough vitamin D_3 for human needs. In the absence of contact between sunlight and skin, dietary supplements become necessary. Fish liver oil, eggs, liver, fortified bread, and milk are the main sources. Deficiency of vitamin D occurs when there is lack of sunlight and inadequate vitamin D in the diet. It may also result from disease or after resection of the small intestine, which may cause malabsorption. In these circumstances, softening of bone (osteomalacia) and rickets occur.

Deficiency of vitamin D

In the jejunum, vitamin D is incorporated along with bile salts and fatty acids into the micelles, and subsequently, as the provitamin D_3, vitamin D is absorbed in the ileum and then passes into the circulation via the portal vein. A specific blood-borne protein, an alpha-1-globulin, carries it to the liver, where the process of chemical change to the active hormone begins by hydroxylation to cholecalciferol. The derivatives are conveyed from the liver to various tissues, including the skin, bone, and parathyroid glands. In the intestine, vitamin D influences the permeability of the brush borders of the enterocytes to calcium.

Intestinal gas. The movement of gas through the intestines produces the gurgling sounds known as borborygmi. In the resting state, there are usually about 200 millilitres of gas in the gastrointestinal tract. Its composition varies: between 20 and 90 percent is nitrogen, up to 10 percent is oxygen, up to 50 percent is hydrogen, up to 10 percent methane, and between 10 and 30 percent is carbon dioxide. Most of the air we swallow, while talking and eating in particular, is either regurgitated (as in belching) or is absorbed in the stomach.

Although some of the carbon dioxide in the small intestine is due to the interaction of hydrogen ions of gastric acid with bicarbonate, some is generated in the jejunum by the degradation of dietary triglycerides to fatty acids. High levels of carbon dioxide in rectal flatus reflect bacterial activity in the colon. Methane cannot be produced by any human cell and is entirely the result of bacteria acting on fermentable dietary residues in the colon, although there appears to be a familial factor involved in this as not everyone can generate methane. How much of this is genetic and how much is environmental is not known. In the colon, bacterial production of hydrogen is markedly elevated when the diet contains an excess of vegetable saccharides. This is particularly noticeable after a meal of baked beans, for example. Gas is more often responsible for the buoyancy of stools than is excessive residual fat in malabsorption states.

The gradient between the partial pressures (or the pressure exerted by each gas in a mixture of gases) of particular gases in the intestinal lumen and the partial pressures of gases in the circulating blood determines the direction of movement of gases. Thus, because oxygen tends to be in low pressure in the colon, it diffuses out from the blood into the intestine. The diffusion of nitrogen from the blood into the intestine occurs because a gradient is established by carbon dioxide, methane, and hydrogen that result from the metabolic activities of the commensal bacteria; the partial pressure contributed by nitrogen in the colon is lowered, stimulating nitrogen to enter the intestine from the blood.

Direction of gas movement

Anxiety induces frequent swallowing of air with consequent belching or increased rectal flatus. The action of sighing sucks air into the stomach as it creates a negative pressure (or one that is less than that of the atmosphere) in the thoracic cage. Fast eaters develop uncomfortable gaseous distension of the abdomen, and swallowed air may reach the anus within 15 to 30 minutes. Rectal gas in air-swallowers contains a high proportion of nitrogen, whereas in excess flatus due to bacterial metabolism of food residues, the gas has a high content of hydrogen and carbon dioxide.

In areas where lactase, the enzyme that breaks down lactose (milk sugar), is missing from the group of disaccharidases of the small intestine, lactose passes into the colon undigested. In the lactase-deficient subject, the unhydrolyzed lactose enters the colon, where the amount of lactose normally present in a glass of milk is capable of liberating, after bacterial fermentation, the equivalent of two cups (500 millilitres) of gas (hydrogen), about 15 percent of which diffuses back into the blood, with the rest passing as flatus.

Hydrogen generated in the colon is partly absorbed, passes in the circulating blood to the lungs, and diffuses into the respiratory passages, where its presence can be easily determined. The time taken for hydrogen to appear in the breath after ingestion of a standard load of glucose or lactose is used to determine whether the upper area of the gastrointestinal tract is colonized by bacteria. Hydrogen that appears within 30 minutes of the ingestion of the sugar load suggests heavy colonization of the small intestine.

Hormones of the gastrointestinal tract. Control of the activity of the specialized cells in the digestive system that are concerned with motor and secretory functions depends upon signals received at their cell membranes. These sig-

Hormones of the Digestive Tract*

notation of cell type	product	location
A	glucagon	pancreas
B	insulin	pancreas
D	somatostatin	body, antrum, duodenum, jejunum, ileum, colon, pancreas
G	gastrin	antrum, duodenum
I	cholecystokinin	duodenum, jejunum, ileum
K	gastric inhibitory peptide	duodenum, jejunum, ileum
L	intestinal glucagon	ileum, colon
Mo	motilin	duodenum, jejunum
N	neurotensin	ileum
PP	pancreatic polypeptide	pancreas
S	secretin	duodenum, jejunum
VIP	vasoactive intestinal peptide	ileum, pancreas, nerve plexuses
EC	hydroxytryptamine	body, antrum, duodenum, jejunum, ileum, colon, pancreas
ECL	histamine	body, antrum, duodenum, jejunum, ileum, colon, pancreas
—	endorphins	body, antrum, duodenum, jejunum, ileum, colon
—	enkephalins	body, antrum, duodenum, jejunum, ileum, colon
—	substance P	body, antrum, duodenum, jejunum, ileum, colon, nerve plexuses
—	bombesin	body, antrum, duodenum, jejunum, ileum, colon, nerve plexuses

*According to the Santa Monica Classification, 1980.

Messenger molecules

nals originate in either endocrine or nerve cells and are carried to the target cell by amino or peptide "messenger" molecules. When secreted, these substances either diffuse into the tissue spaces around the cells and affect target cells in the vicinity or are taken up in the circulating blood and delivered to target cells some distance away. In both circumstances, the messengers are hormones, but those exerting their effect locally are called paracrine; those exerting their effect at a distance are called endocrine.

Peptides are composed of a number of amino acids strung together in a chain. The amino acids occur in an ordered sequence that is peculiar to each peptide. The biological activity of the peptide (*i.e.*, the ability to stimulate the target cells) may reside in only a fraction of the chain, for example, in a four- or five-amino-acid sequence. In other instances, the entire chain must be intact to achieve this purpose. For example, dispersed throughout the whole gastrointestinal tract are the delta (D) cells (see Table), which produce a hormone known as somatostatin. This hormone has inhibiting effects on the production of acid in the stomach, the motor activity of the intestine, and the release of digestive enzymes from the pancreas. These effects are achieved by local diffusion of somatostatin from the delta (D) cells in the vicinity of the target tissue; there is no evidence that somatostatin circulates in the blood. On the other hand, gastrin, a hormone produced by the granular enterochromaffin (G) cells in the mucosa of the gastric antrum, is secreted into the blood.

Gastrin also exemplifies the biological capability of a fraction of the molecule. These fractions have a molecular structure that fits the receptor site on the membrane of the target cell and therefore can initiate the intracellular events in the production of the acid. The G cells produce a messenger peptide with 34 amino acids in sequence and another with only 17 in the chain. The shorter molecule is more potent. The chain can be cleaved to only four amino acids (the tetrapeptide), however, and providing the sequence remains the same as in the parent molecule and the fragment is the one at the amino terminal of the whole molecule, the cleaved amino acid chain retains biological activity, although it is less potent than the larger molecules of gastrin.

Immunocytochemistry

The identification of the gastrointestinal hormones and their cells of origin has been facilitated by the advent of immunocytochemistry, for both light and electron microscopy. In this process, the peptide under study is injected into an animal. Because this is "foreign" protein, or antigen, it stimulates the formation of specific antibodies that fight a particular peptide. In due course the antibodies are harvested from the animal's blood. The antibodies are then "labeled" by attaching a fluorescent dye (for viewing by ordinary light microscopy), or gold (for viewing by electron microscopy). The tissue under examination is treated with this immunochemical complex, and the antibody seeks out the original peptide, or antigen, latches onto it, and deposits the fluorescein or gold at the site. The cell and its contents are then available for detailed microscopy.

Immunocytochemistry made possible the recognition that G cells of the antrum of the stomach mainly produce the 17-amino-acid variety of gastrin, while those in the duodenum and jejunum mainly produce the 34-amino-acid variety. Through the application of such methods, certain messenger peptides have been found to originate not in endocrine cells but in neural elements within the gastrointestinal tract, to be released during electrical discharge within the nerves. For example, vasoactive intestinal peptide (VIP) released from nerve terminals in the brain also is present in abundance in the nervous structures of the gut, including the submucosal and myenteric nerve plexuses. Occasionally VIP coexists with acetylcholine, the messenger molecule of the autonomic parasympathetic nervous system. The discharge of VIP brings about receptive relaxation of the esophageal and pyloric sphincters, modulates the long peristaltic movements in the intestine, and influences the secretion of electrolytes from the mucosa of the small intestine.

Eighteen different endocrine cells can be identified within the gastrointestinal tract, but it is probable that several of these and their particular peptides are evolutionary vestiges that functioned in other stages of human development, while others may represent different stages of maturation of the same endocrine cell.

Peptides that bind with target cell receptors and stimulate the cell to react are known as agonists. Others that fit the receptor but do not initiate intracellular events are known as antagonists. The ability of antagonists to occupy receptors and thereby deny access to an agonist is the basis of the treatment of peptic ulcer disease with histamine (H_2) receptor blockers. By occupying the receptors on the parietal cells, antagonists deny histamine the opportunity to initiate the production of hydrochloric acid.

Antagonists

It should be noted that similar events stimulate or suppress the production of the messenger peptides in their endocrine or neural cell of origin. For example, the discharge of granules of gastrin from the G cells occurs when a meal is taken. While the concentration of hydrogen ions (the acidity) remains low because of the buffering effect of the food, the release of gastrin continues. As digestion proceeds and the stomach begins to empty, however, the acidity increases because of the diminishing neutralizing effect of the food. When the contents of the stomach in contact with the mucosa of the antrum reach a certain level of acidity (pH of 2.5 or less), the release of gastrin stops. Failure of this mechanism leads to inappropriate secretion of acid when the stomach is empty of food and may cause peptic ulcers in the duodenum. Some endocrine cells have microvilli on their surface that project into the lumen of the gland or into the main channel of the stomach or intestine. These cells probably have an ability to "sample" continuously the lumenal contents in their vicinity. This could be the basis of the information pertinent to the behaviour of the endocrine cell.

When production and secretion of a peptide hormone is excessive, it induces an increase in the number of the target cells and may increase the size of the individual cells. This is known as trophism and is similar to the increase in size of skeletal muscle in response to appropriate exercise (work hypertrophy). Such trophism is observed in certain disease states that involve the gastrointestinal hormones. Thus, when gastrin is secreted into the blood by a tumour of G cells (gastrinoma) of the pancreas, it is a continuous process because there is no mechanism at that site to inhibit the secretion: this brings about a massive increase in the number of parietal cells in the stomach and an overproduction of acid. This overwhelms the defenses of the mucosa of the upper gastrointestinal tract against autodigestion and results in intractable and complicated peptic ulceration.

Characteristics of gastrointestinal hormones

The individual endocrine cells and their products are listed in the Table. The characteristics of the gastrointestinal hormones are summarized below.

Insulin. Insulin is secreted by the beta (B) cells of the pancreas in response to a rise in plasma glucose concentration and a fall in glucagon level. It disperses carbohydrate (glucose) to stores in muscle and adipose (fatty) tissue. Insulin is used in the treatment of diabetes mellitus.

Glucagon. A hormone that is elaborated by alpha (A) cells, which are widely distributed in the alimentary tract, glucagon may stimulate the secretion of water and electrolytes by the mucosa of the small intestine. Otherwise, glucagon is an inhibiting hormone. It inhibits the production of gastric and pancreatic secretions—in particular the secretion of insulin—and the contractions of smooth muscle. It is sometimes used in radiology and endoscopy to improve visualization of the duodenum. Glucagon also is used in the treatment of conditions in which the level of sugar in the blood is lowered.

Somatostatin. Somatostatin is a peptide secreted by the delta (D) cells in response to eating, especially when fat enters the duodenum. It is an inhibitory modulator of the secretion of acid and pepsin and the release of gastrin, insulin, and other intestinal hormones. It inhibits motility of the gallbladder and intestines and suppresses the secretion of lipase by the pancreas.

Serotonin. Serotonin, or 5-hydroxytryptamine, is an amine that is formed from amino acid 5-hydroxytrytophan, in the enterochromaffin cells (EC) and in other similar

cells called enterochromaffin-like cells (ECL). These cells also secrete histamine and kinins, which likewise have important messenger functions in glandular secretions and on blood vessels. Serotonin acts in paracrine fashion. Both EC and ECL cells are widely distributed in the gastrointestinal tract.

Cholecystokinin. A peptide secreted by the I cells in response to the emptying of the stomach contents into the duodenum, cholecystokinin causes contraction of the gallbladder with emptying of its contents, relaxation of the sphincter closing the end of the bile duct, and stimulation of the production of enzymes by the pancreas. Cholecystokinin increases intestinal peristalsis, and it is used in radiological examination of the gallbladder and in tests of pancreatic function.

Uses of cholecystokinin

Gastric inhibitory peptide. Secreted by the K cells, gastric inhibitory peptide enhances insulin production in response to high concentration of blood sugar, and it inhibits the absorption of water and electrolytes in the small intestine. The cell numbers are increased in persons with duodenal ulcer, chronic inflammation of the pancreas, and diabetes resulting from obesity.

Intestinal glucagon. Secreted by the L cells in response to the presence of carbohydrate and triglycerides in the small intestine, intestinal glucagon (enteroglucagon) modulates intestinal motility and has strong trophic influence on mucosal structures.

Motilin. A high level of motilin in the blood stimulates the contraction of the fundus and antrum and decreases gastric emptying. It contracts the gallbladder and increases the squeeze pressure of the lower esophageal sphincter. Motilin is secreted between meals.

Neurotensin. Secreted by the N cells of the ileum in response to fat in the small intestine, neurotensin modulates motility, relaxes the lower esophageal sphincter, and blocks the stimulation of acid and pepsin secretion by the vagus nerve.

Pancreatic polypeptide. Special endocrine cells, "PP" cells, secrete pancreatic polypeptide in response to protein meals. Their function is intimately related to vagal and cholinergic activity. The level of pancreatic polypeptide is frequently raised in diabetes.

Secretin. Secreted by the S cells of the duodenum in response to meals and to the presence of acid in the duodenum, secretin stimulates the production of bicarbonate by the pancreas.

Vasoactive intestinal peptide. Secreted locally by endocrine cells or nerve endings, vasoactive intestinal peptide is located almost exclusively in nerves distributed throughout the gastrointestinal tract. It inhibits the release of gastrin and the secretion of acid and is a mild stimulant of bicarbonate secretion from the pancreas and a powerful stimulant of the secretion of water and electrolytes by the small and large intestines. It relaxes the sphincters and slows intestinal transit time. There is a further group of peptide messengers that is found in quantity within the brain and in the nerves of the gastrointestinal tract. These include substance P, endorphins, enkephalins, and bombesin.

Substance P. Present in significant amounts in the vagus nerves and the myenteric plexus, substance P stimulates saliva production, contraction of smooth muscle cells, and inflammatory responses in tissues, but it is uncertain whether it is other than an evolutionary vestige.

Endorphins and enkephalins. Endorphins and enkephalins, each comprising five amino acids in the molecule, are present in the vagus nerves and the myenteric plexus. They have the properties of opiate (opium-derived) substances such as morphine; they bind to the same receptors and are neutralized by the opiate antagonist naloxone. There is no evidence that endorphins and enkephalins are circulating hormones, but the enkephalins may have a physiological paracrine role in modulating smooth muscle activity in the gastrointestinal tract, and endorphins may serve in modulating the release of other peptides from endocrine cells in the digestive system.

Properties of endorphins and enkephalins

Bombesin. A peptide that is found in the intrinsic nerves of the gastrointestinal tract, bombesin stimulates the release of gastrin and pancreatic enzymes and causes contraction of the gallbladder. These functions may be secondary, however, to the release of cholecystokinin, a hormone secreted by the mucosa of the intestine that has similar effects. It is uncertain if bombesin has a physiological role in humans or if it is an evolutionary vestige.

THE GASTROINTESTINAL TRACT AS AN ORGAN OF IMMUNITY

The body is continuously exposed to damage by viruses, bacteria, and parasites; ingested toxins and chemicals, including drugs and food additives; and foreign protein of plant origin. These insults are received by the skin, the respiratory system, and the digestive system, which constitute the interface between the "perfect" and sterile body interior and the environment.

The defense of the body is vested largely in the lymphatic system and the lymphocytes in it. A substantial part of the gastrointestinal tract is occupied by lymphoid tissue, which can be divided into three sectors. The first is represented by the pharyngeal tonsils, the appendix, and the large conglomerates known as Peyer's patches sited at intervals throughout the small intestine. The second sector includes the lymphocytes and plasma cells that populate the basement membrane (lamina propria), the loose connective tissue area above the supporting tissue of the mucosal lining extending into the villi. The third sector comprises lymphocytes, which lie between the epithelial cells in the mucosa. The basis of defense of the gastrointestinal tract is the interaction between these cells of the lymphatic system and the threatening agent.

Lymphocytes are of two types, B and T, according to whether they originate in the bone marrow (B) or in the thymus gland in the chest (T). On leaving their tissue of origin, both types end up in the peripheral lymphoid structures. These include the peripheral lymph glands, the spleen, the lymph nodes in the mesentery of the intestine, the Peyer's patches, and the spaces between the epithelial cells of the mucosa.

Lymphocyte classes

Lymphocytes are described as immature until they come into contact with antigens. If this foreign material is recognized as such by T cells (T lymphocytes), the cells undergo a process of maturation. They proliferate and divide into subclasses. The first subclass comprises the "helper" T cells, which are mediators of immune function. The second class consists of "suppressor" T cells, which modulate and control immune responses. The third class comprises the "killer" T cells, which are cytotoxic (*i.e.,* they are able to destroy other cells). Most of the lymphocytes lying between the epithelial cells of the mucosa are killer T cells.

When B cells (B lymphocytes) recognize antigen, they also mature, changing to the form known as plasma cells. These cells elaborate a highly specialized protein material, immunoglobulin (Ig), which constitutes antibodies. There are five varieties of immunoglobulin, IgA, IgM, IgG, IgD, and IgE. The cells in the spaces of the basement membrane are mainly B cells and plasma cells. Another group of specialized cells are known as M cells. These are stretched over and around ordinary epithelial cells of the mucosa. The M cells have a function in transportation: they package antigenic material into vesicles and move it through the cell and into the surrounding spaces.

Lymphocytes of the Peyer's patches undergo migration. They pass through lymph vessels to the nodes in the mesentery and then to the thoracic duct. This is the collecting channel in the abdomen, which passes up through the thorax to drain into the venous system at the junction of the left internal jugular and left subclavian veins. The various ramifications of the abdominal lymphatics all drain into the thoracic duct. From there, the lymphocytes are carried back to the intestine as well as being dispersed to other organs. It is these migrated lymphocytes that come to populate the basement membrane and to occupy the spaces between epithelial cells.

Most cells in the mesenteric nodes and the basement membrane are plasma cells that produce immunoglobulin of class IgA, while IgM and, to a lesser extent, IgE are produced by other cells, and IgG is formed by cells in the spleen and peripheral lymph nodes. The IgA of plasma cells is secreted into the lumen of the intestine, where it

Immunoglobulin classes

is known as "secretory IgA" and has a different molecular structure from that of the IgA circulating in the blood. When secreted, it is accompanied by a glycoprotein that is produced by the epithelial cells of the mucosa. This substance, when attached to the IgA molecule, protects it from digestion by protein-splitting enzymes. This IgA complex can adhere to virus and bacteria, interfering with their growth and diminishing their power to invade tissue. It is also capable of rendering toxic substances harmless.

Formed by B cells, IgE coats the surface of mast cells, which are specially adapted to deal with the allergic challenge posed by parasites and worms.

The newborn infant is protected by already-matured immunoglobulin with which the colostrum is richly endowed. Colostrum is the initial secretion of the lactating breast. As time passes, the gastrointestinal tract of the infant is increasingly exposed to various insults, and the lymphocytes and other cells of the immune system become adapted to deal with these. In this way, the body also develops a tolerance to potentially offending substances. If invasion of tissue occurs despite these various defenses, then a generalized systemic immune reaction is marshaled. Some of the features of this reaction, such as fever and a massive increase in the white blood cells, are the evidence of illness. For detailed treatment of the immune system, see IMMUNITY.

(W.S.)

Disorders and diseases of the digestive system

IMMUNE-RELATED DISORDERS

Some individuals fail to produce adequate amounts of certain classes of immunoglobulin. Consequently, they suffer from overgrowth of bacteria in the small intestine and from protozoal and other parasitic infections. Because this often leads to massive accumulations of lymphocytes and plasma cells, the lining of the intestine may become studded with nodules that may interfere with the absorptive functions of the small intestine.

IgA deficiency. Immunoglobulin A (IgA) deficiency is a selective disorder that affects about one out of every 500 infants and arises out of a genetically endowed defect. It results in the inadequate production of IgA. These individuals are prone to allergic reactions to a variety of proteins, such as that in cow's milk. They also have a low resistance to infection by fungi.

Genetically endowed defect

There are a number of uncommon conditions that result from defects in the immune system. They are usually evident as chronic diarrhea, malabsorption, or a relative failure of mucosal cell repair and replacement, resulting in a tendency for the intestinal mucosa to atrophy. These include a disorder known as alpha-chain disease, in which the molecular structure of the IgA is faulty and the amino acid sequence in the chain is disturbed. Besides diarrhea and "opportunistic" infections with viruses, bacteria, or fungi (so called because they take advantage of the reduced immune capability of the individual) there is an increased risk of ultimately developing malignant disease.

Failure of this kind in the immune system may have resulted from other diseases of the intestinal tract. For example, a number of diseases can so damage the gastrointestinal lining that protein leaks out and is lost in the feces. This state is designated by the general term protein-losing enteropathy. In this condition, lymphocytes and immunoglobulin are lost as well.

Organ or tissue rejection after transplantation occurs when the graft and host tissues show a disparity in the tissue markers of the sixth chromosome (known as the HLA histocompatability complex). In the gastrointestinal tract this reaction leads to severe destruction and subsequent atrophy of the mucosa, resulting in malabsorption.

Another failure of the immune system is due to a deficiency in the ability of certain white blood cells, the neutrophils (leukocytes), to kill the viruses and bacteria that they ingest. As a result, mucosal destruction, severe diarrhea, and secondary malabsorption occur with chronic inflammation in the mucosa. The condition is known as chronic granulomatous disease of childhood, "granulomatous" because of the large conglomerates of lymphocytes and giant multinuclear cells that develop in the mucosa.

Chronic granulomatous disease

Most of the deficiency states of the body's immune system have a heightened risk of subsequent malignant disease, especially of the lymphomas, which involve the lymphoid tissues.

Crohn's disease. A chronic inflammatory and ulcerative disease of the gastrointestinal tract, named after one of the physicians who first described it, Crohn's disease has many features of abnormal immune reactions. These include the production of antibodies to colonic bacteria and to milk proteins. Various complications affecting the liver, skin, and joints indicate damage due to the deposition of immune complexes, which arise from the interaction between the antigen and the antibody. The combined antigen–antibody complex circulates in the blood, and those that are not promptly removed by the liver or other components of the reticuloendothelial system (a defensive system of the body comprising phagocitizing cells) may be deposited in the linings of the blood vessels and in the skin, the joints, the kidneys, and elsewhere, commencing a disease process.

MOUTH AND ORAL CAVITY

Besides local disease, features characteristic of systemic disorders are often present on the mouth and in the oral cavity. The lips may be fissured and eroded at the corners in riboflavin deficiency (angular chelitis). Multiple brown freckles on the lips associated with polyps in the small intestine is characteristic of Peutz-Jegher's syndrome. Spider nevi, which are prominent in chronic liver disease, are not confined to the face or congregated on the lip margins. Aggregates of small yellow spots on the buccal mucosa and the mucosa behind the lips indicate Fordyce's "disease." The spots are due to the presence of enlarged sebaceous glands just below the mucosal surface.

The most common mouth ulcers are due to aphthous stomatitis. These affect one out of every five Caucasians. The spectrum of this condition ranges from one or two small painful vesicles rupturing to form round or oval ulcers, occurring once or twice a year and lasting seven to 10 days, to deep ulcers of one centimetre or more in diameter. The ulcers are frequently multiple, occur anywhere in the mouth (on the tongue or the palate), and may persist for months at a time. The disability ranges from a mild local irritation to severe distressing pain that prevents talking and eating. Scarring can be seen at the sites of previous ulcers. Aphthous ulceration is sometimes associated with psychological stress, but it may also be a reflection of an underlying malabsorptive disease such as celiac disease. Treatment is directed to the predisposing cause. Local anesthetic agents and analgesics may permit talking and eating. Topical and systemic corticosteroids are the most effective treatment. In a more serious condition, Behçet's syndrome, similar ulcers occur in the mouth and on the genitalia, and the eyes are involved.

Aphthous ulcers

Discoloration of the tongue, commonly white, is due to deposits of epithelial debris, effete (or worn out) bacteria, and food. It occurs in circumstances in which there is reduced saliva production. This may be acute, as in fevers, when the body temperature is high and loss of water through the skin is excessive. The reduced saliva flow in fever is a conservation phenomenon, akin to oliguria (reduced urine losses) and constipation. The discoloration becomes chronic following atrophy of the glands and in the absence of good oral hygiene. If the person is a heavy smoker, the deposit is coloured brown. Black discoloration of the tongue with the formation in the centre of a dense pellicle of fur (black hairy tongue) may be due to a fungus with pigmented filaments. Occasionally it simply represents excessive elongation of the filiform papillae. It may be due to sucking licorice candy.

Tongue discoloration

A bald tongue, with a smooth surface due to complete atrophy of the papillae, is seen in severe iron-deficiency anemia, pernicious anemia, and pellagra, a disorder of skin and mucous membranes due to niacin deficiency. This condition occurs in cereal eaters, usually of corn, when the cereal is contaminated or has a low content of the vitamin that leads to an imbalance in the amino acids derived from food. The condition is endemic in underdeveloped countries in which there are periods of famine.

A deeply fissured tongue (scrotal tongue) may be due to a congenital variation in the supporting tissue of the tongue, but it can be acquired. There is a mild degree of inflammation in the fissures, which causes a slight burning discomfort.

Geographic "tongue," or migrating exfoliative glossitis, describes areas of denudation of the surface of the tongue of various shapes and sizes. These gradually become re-epithelialized with regrowth of the filiform papillae, only for the inflammatory process to begin elsewhere in the tongue. Thus, the picture changes with time as the bald zones move around the tongue. These changes usually give rise to no symptoms or, at the most, to a mild burning sensation. The cause is unknown, and the condition may persist for years.

Trench mouth

Vincent's disease (trench mouth) is an ulcerating, necrotizing infection of the gingiva (gums) notable for the spontaneous bleeding from affected areas and the foul odour of the breath arising from the gangrenous tissue. It is endemic in countries where there is severe malnutrition and poor oral hygiene. The infection probably involves several organisms, including spirochetes and fusiform bacilli. It is uncertain if in the developed countries it is transmitted by the exchange of saliva in kissing, but its epidemic increase in wartime and its frequency in the promiscuous suggest this. Vincent's disease responds to antibiotics followed by trimming of the gum margins to eliminate subgingival pockets.

Malignant disease, or cancer, of the mouth is sometimes caused by chronic thermal irritation in heavy smokers and is often preceded by leukoplakia (plaque-like patches arising on the mucous membranes of the cheeks, gum, or tongue). Similarly, cancer of the mouth can be caused by the habit of keeping tobacco or a package of intensely "hot" spices in the space between the cheek and the teeth. These cancers arise from the squamous cells that line the oral mucosa. Cancers of the salivary glands and of the mucous membranes of the cheeks cause pain, bleeding, or difficulty in swallowing. The lymphomas and other tumours of lymphoid origin may first appear in the tonsillar or pharyngeal lymph nodes. Cancer of the tongue and of the bony structures of the hard palate or sinuses may project into the mouth or may burrow deep into the surrounding tissues.

Dental caries. Dental caries are due to the destruction of the dental enamel and underlying tissues by organic acids. These acids are formed by bacteria growing in debris and food accumulated in pockets between the base of the teeth and the gum margins. This periodontal infection ultimately leads to the invasion of the dental pulp, and the involvement of the nerve in the inflammation is the cause of toothache. An abscess may form at the apex of the tooth and extend into the jawbone, causing osteomyelitis (inflammation of the bone), or into the soft tissues around the roots of the teeth, causing cellulitis (inflammation of the soft tissues). Halitosis is due to the rotting debris in the pockets under the gum margins. In due course the teeth loosen and fall out or need to be extracted. Poor oral hygiene is the underlying predisposing circumstance. Malnutrition due to poverty, alcoholism, and malabsorption of vitamin D (rickets) or of proteins (as in celiac disease), initiate or aggravate caries. The resistance of the dental enamel to damage by organic acids is increased by fluoride, and in many countries this is incorporated into the toothpaste formula and is added to the water supplied to homes. In areas where these steps have been taken, the incidence of caries has dropped by more than 50 percent.

Pharyngitis. Inflammation of the posterior wall of the mouth and of the tonsils and adjoining tissue on each side of the oropharynx is very common, especially in young persons. Such infections are due to bacteria of the streptococcal and staphylococcal species, but it should be emphasized that just as many diffuse inflammatory reactions in these tissues result from viral infections as from invasion by pyogenic (pus-forming) bacteria. In viral pharyngitis the tissue is usually less violently red and swollen than is true of streptococcal pharyngitis and it is less often covered by a whitish exudate. Other tonsillar tissue in the upper part of the pharynx and at the root of the tongue may be similarly involved. In diphtheritic pharyngitis, the membranous exudate is more diffuse than in other types of pharyngitis, it is tougher, and it extends over a much larger part of the mucous membrane of the mouth and nose. One of the complications of any tonsillitis or pharyngitis may be a peritonsillar abscess adjacent to one tonsil; this appears as an extremely painful bulging of the mucosa in the area. Surgical evacuation is usually required.

Causes of pharyngeal infections

Inborn defects. The most important of the inborn defects is harelip, which involves failures in fusion of the palate (the bones and soft tissues of the roof of the mouth), thus impairing the ability to produce a closed, high-pressure cavity behind the lips and teeth. Other disorders are related to defective apposition of the teeth and the jaws, resulting in inefficient mastication, and to the absence of one or more of the salivary glands, which may lessen the amount and quality of saliva that they produce. Neurological defects that provide inadequate motor power to the muscles of the tongue and the pharynx can seriously impair mastication and even swallowing. Sensory-innervation defects may not allow the usual reflexes to mesh smoothly, or they may permit harmful ingestants to pass by undetected.

SALIVARY GLANDS

Hyposecretion of saliva

The secretion of saliva is markedly diminished in states of anxiety and anxiety depression. The consequent dry mouth interferes with speech, which becomes thick and indistinct. In the absence of the cleansing action of the saliva, food debris persists in the mouth and stagnates, especially around the base of the teeth. The debris is colonized by bacteria and causes foul breath (halitosis). In the absence of saliva, swallowing is impeded by the lack of lubrication for the mastication of food that is necessary to form a bolus. The condition is aggravated in states of anxiety depression when drugs that have an anticholinergic-like activity (such as amitriptyline) are prescribed, because they further depress the production of saliva. The salivary glands are severely damaged and atrophy in a number of autoimmune disorders such as Sjögren's disease and systemic lupus erythematosus. The damage is done partly by the formation of immune complexes (antigen–antibody associations), which are precipitated in the gland and initiate the destruction. In these circumstances, the loss of saliva is permanent. Some symptomatic relief is obtained by the use of "artificial saliva," methylcellulose mouthwashes containing herbal oils such as peppermint. As some of the salivary glands retain their function, they may be stimulated by chewing gum and by a parasympathomimetic agent such as bethanecol. The production of saliva may be impaired by infiltration of the salivary glands by pathological lymphocytes, such as in leukemias and lymphomas. In the early stages of these diseases, the glands swell and become painful.

Excessive production of saliva may be apparent in conditions interfering with swallowing, as in parkinsonism due to disease of the brain stem and basal ganglia of the brain, or in pseudobulbar paralysis from blockage of small arteries to the midbrain regions. Both conditions bring about drooling. True salivary hypersecretion is seen in poisoning due to lead or mercury used in certain industrial processes, and as a secondary response to painful conditions in the mouth, such as aphthous stomatitis (certain ulcers of the oral mucosa) and advanced dental caries.

Acute and painful swelling of a salivary gland develops when salivary secretion is stimulated by the sight, smell, and taste of food but a salivary duct is obstructed. The swelling and pain that develop because saliva is prohibited from flowing through the duct subside between meals. The diagnosis can be confirmed by X ray. Persistent swellings may be due to infiltration by benign or malignant tumours or to infiltration by the abnormal white blood cells, as in leukemia. The most common cause of acute salivary swelling is mumps.

ESOPHAGUS

Dysphagia

Difficulty in swallowing (dysphagia) may be the only symptom of a disorder of the esophagus. Sometimes difficulty in swallowing is accompanied by pain (odynophagia), or

pain may occur spontaneously without swallowing being involved. The esophagus does nothing to alter the physical or chemical composition of the material it receives, and it is poorly equipped to reject materials that have gotten past the intricate sensors of the mouth and throat. Consequently, it is peculiarly vulnerable to mucosal injury from ingestants, as well as to materials that reflux into its lower segment from the stomach. Although its muscle coats are thick, it is not protected with a covering of serous membrane, as are neighbouring organs in the chest.

Congenital disorders. Inborn defects of the esophagus are most often seen in infancy, primarily as a failure to develop normal passageways. An opening often occurs between the esophagus and the trachea. Babies born with these openings cannot survive without early surgery. The lower end of the esophagus is subject to various developmental anomalies that shorten the organ so that the stomach is pulled up into the thoracic cavity. Anomalies of the diaphragm may contribute to a similiar outcome.

Inflammatory disorders of the esophagus result from a variety of causes; for example, ingestion of noxious materials as in lye or acid burns, lodgment of foreign bodies, or a complex of events associated with reflux of gastric contents from the stomach into the lower esophagus. All types of trauma produce damage to the mucous membrane of the esophagus. Inflammation resulting from surface injury by caustics is called corrosive esophagitis. When the problem is associated with reflux, the term peptic esophagitis is applied to inflammation, which involves both the mucous membrane and the submucosal layer in a generally mild process. A number of other diseases may cause inflammation of the esophagus; *e.g.*, scleroderma, a disease in which the smooth muscle of the organ degenerates and is eventually replaced by fibrous tissue; and generalized moniliasis, in which the esophagus is often involved in a septic process characterized by many small abscesses and ulcerations throughout its entire length.

Peptic esophagitis

Strictures. It is characteristic of all fibrous (scar) tissue that it contracts over time. Consequently, when fibrous tissue is laid down around a tube, as in the esophagus, in response to inflammation, the contracting scar narrows the lumen, causing a stricture, and may come to obstruct it totally. Strictures are readily diagnosed by X ray and by viewing directly through an esophagoscope.

Dysphagia. Most individuals can locate the site of the dysphagia (difficulty in swallowing) and the distribution of the pain with accuracy. A sense of food sticking or of pain on swallowing, however, may be felt to be in the throat or upper sternum when the obstruction or disease is in fact at the lower end of the esophagus. The sensation of a "lump in the throat" that is not connected with eating or swallowing is called "globus hystericus." This is an expression of a psychoneurosis and is accompanied by other symptoms indicative of this. Globus hystericus is managed by explanation and treatment of the underlying nervous disorder.

The neural arc of swallowing involves the medulla of the brain stem, the vagus (10th cranial) nerves, and the glossopharyngeal, trigeminal, and facial nerves. Consequently, dysphagia may result from interference with the function of any part of this pathway. Thus, it occurs commonly, but usually transiently, in strokes. Dysphagia may be prominent in degenerative disease of the central nervous system, especially of the ganglia at the base of the brain. In these circumstances, it is the behaviour of the smooth muscle of the pharynx and the upper esophageal sphincter that is disturbed.

Pain. The nerve fibres conveying the sense of pain from the esophagus pass through the sympathetic system in the same spinal cord segments as those that convey pain sensations from the muscle and tissue coverings of the heart. Episodes of pain arising from the esophagus as a result of muscle spasm or transient obstruction by a medicine tablet or other object may be experienced in the chest and posterior thorax and radiate to the arms. It thus mimics angina (pain) of cardiac origin (pseudoangina). The pain due to transient obstruction may be felt not only in the chest but also, through radiation to the back, between the shoulder blades. Because this is very similar to pain from gallstones, attacks lasting 10 to 30 minutes mimic that condition.

Pseudo-angina

In middle-aged and elderly persons, spontaneous and diffuse spasm of the smooth muscle of the esophagus causes considerable discomfort as well as episodes of dysphagia. Alternative names for the condition are "corkscrew" esophagus and diffuse spasm of the esophagus. The appearance of the esophagus seen on an X-ray screen while a barium bolus is swallowed resembles that of the outline of a corkscrew because of the multiple synchronous contractions at different levels of the spirally arranged smooth muscle.

Drugs that relieve cardiac angina may also relieve the pain of esophageal spasm, especially nifedipine. In persons over 50 years of age, the sensation of food "sticking" is more often caused by a disease process, frequently a tumour, involving the wall of the esophagus and providing a mechanical rather than a functional obstacle to the passage of food.

Motility. Disorders of the motility of the esophagus tend to be either precipitated or aggravated at times of nervous stress. Eating rapidly is another trigger, as this demands more precise and rapid changes in muscle activity than eating slowly. An interesting disease of the esophagus is achalasia, formerly called cardiospasm. In this disorder, a primary disturbance in the peristaltic action of the esophagus results in failure to empty the organ of its ingested and swallowed contents. The lower sphincteric portion of the esophagus does not receive its customary signal to relax and, over the months and years, may become hypertonic, resisting stretching. A vicious cycle is thus set up in which the main portion of the esophagus slowly becomes distended, holding a column of fluid and food that it cannot propel downward to a lower exit valve mechanism that stays closed because of a failure in its information-feedback system. In most persons with this disorder, there is a shortage or disease of ganglion cells of the intramuscular plexus (Auerbach's plexus), or a disease of the network of nerves within the muscles of the esophagus, so that coordinated peristalsis becomes impossible. In Chagas' disease, the infecting trypanosomes invade the neural tissue and directly destroy the ganglion cells. These organisms are not present in the temperate zones of the world, however, and the reason for ganglion cell degeneration in achalasia is generally unknown. Effective treatment is achieved by destroying the ability of the lower esophageal sphincter of the esophagus to contract. This may be done by forcible dilatation, using a balloon, of the esophagus in the area that is tonically contracted. The objective is to rupture the circular muscle at that site, and this is generally achieved with one or two dilatations. If this fails to overcome the contraction or if the contraction recurs, the surgical cure involves opening the abdomen and cutting through the circular muscles from the outside of the esophagus. The disadvantage of both methods of treatment is that the antireflux mechanism is thereby destroyed. Consequently, if precautions are not taken, the individual may lose the symptoms and risks of achalasia but may develop the symptoms and signs of reflux peptic esophagitis.

Onset of disorders of motility

Gastroesophageal reflux. In healthy individuals, reflux of gastric contents into the esophagus occurs occasionally. This causes the burning sensation behind the sternum that is known as heartburn. It may be accompanied by regurgitation, some of the refluxed material reaching the pharynx where it also may be felt as a burning sensation. Reflux is most likely to occur after large meals, especially if physical activity, including bending, stooping, or lifting, is involved. In these circumstances, the esophagus responds with peristaltic waves that sweep the gastric contents back into the stomach, with rapid relief of the heartburn.

Persistent reflux symptoms are invariably due to inadequate functioning of the anatomical components, such as the lower esophageal sphincter, which keep the contents of the stomach below the diaphragm. The disorder is most commonly due to obesity. Excessive fat on the trunk is almost always accompanied by large deposits of fat within the abdomen, especially in the mesentery (the curtain-like structure on which most of the intestine is hung). Consequently, when intra-abdominal pressure is increased, such

Mechanism of reflux

as in physical activity, there is insufficient room within the abdomen to accommodate the displacement of the organs, and the resulting pressure forces the stomach upward. The weak point is the centre of the diaphragm at the opening (hiatus) through which the esophagus passes to join the stomach. The upper pole of the stomach is pushed through the hiatus, and the distortion of the anatomical relations brings about impaired functioning of the anti-reflux mechanisms. In the early stages the stomach may slide back into the abdomen when the increase in the intra-abdominal pressure eases, but eventually, if the circumstances are unchanged, the upper part remains above the diaphragm. A common contributory cause in women is pregnancy. As the uterus containing the developing fetus comes to occupy a large part of the abdomen, the effect is the same as in obesity. Because only gravity keeps the gastric contents within that organ, once a hernia is established, the reflux and the symptoms from it promptly occur when the individual lies down. Persisting reflux of gastric contents with acid and digesting enzymes leads to chemical inflammation of the lining of the esophagus and ultimately to (peptic) ulceration. If inadequately treated, the process leads to submucosal fibrosis and stricturing, and, besides the symptoms of heartburn and regurgitation, the patient experiences pain on eating and swallowing.

Treatment of reflux

The treatment of peptic esophagitis includes eliminating obesity, avoiding activity that increases intra-abdominal pressure, and raising the head of the bed high enough to discourage nocturnal gastroesophageal reflux. Mixtures of alkali and alginic acid derivatives that are viscous and adhere to the inflamed area of mucosa are effective and are aided by taking agents that reduce the secretion of acid by the stomach, such as those blocking the receptors on the surface of the parietal cells. If a stricture has formed, it can be dilated easily. If the disability is not overcome with these conservative measures, surgical repair is undertaken through either the chest or the abdomen. When feasible, the vagus nerves may be cut to stop acid secretion permanently.

In some individuals with severe reflux esophagitis, the repair of the damaged lining of the esophagus after conservative treatment is brought about by relining the esophagus with columnar cells. These cells are similar to those lining the upper part of the stomach and are not the usual squamous cells that line the esophageal mucosa. In some persons in whom this transformation occurs, a carcinoma develops some 10 to 20 years later. The decision as to the treatment of a hiatus hernia by conservative means or by surgery is influenced by such factors as the age of the patient, his occupation, and the likelihood of compliance with a strict regimen.

There is a much less common form of hiatus hernia, called paraesophageal, in which the greater curvature of the stomach is pushed up into the thorax while the esophagogastric junction remains intact below the diaphragm. Such individuals experience dysphagia caused by compression of the lower esophagus by the part of the stomach that has rolled up against it. This rarer form of hernia is more dangerous, often being complicated by hemorrhage or ulceration, and requires relief by surgery.

Diverticula. Pouches in the walls of the structures in the digestive system that occur wherever weak spots exist between adjacent muscle layers are called diverticula. In the upper esophagus, these may occur in the area where the striated constrictor muscles of the pharynx merge with the smooth muscle of the esophagus just below the larynx. Some males over 50 years of age show protrusion of a small sac of pharyngeal mucous membrane through the space between these muscles. As aging continues, or if there is motor disturbance in the area, this sac may become distended and actually fill with food or saliva. It usually projects to the left of the midline, and its presence may become known by the bubbling and crunching sounds produced during eating. Often the patient can feel it in the left side of the neck as a lump, which can be reduced by pressure of the finger. Sometimes the sac may get so large that it compresses the esophagus adjacent to it, producing a true obstruction. The treatment is by surgery. Small diverticula just above the diaphragm sometimes are found after the introduction of surgical instruments into the esophagus.

Pharyngeal diverticulum

A serious injury to the esophagus is spontaneous rupture. It can occur in patients who have been vomiting or retching and in debilitated elderly persons with chronic lung disease. Emergency surgical repair of the perforation is required. A rupture of this type confined to the mucosa only at the junction of the linings of the esophagus and stomach is called a Mallory-Weiss lesion. At this site, the mucosa is firmly tethered to the underlying structures and, when repeated retching occurs, this part of the lining is unable to slide and suffers a tear. This leads to immediate pain beneath the lower end of the sternum and bleeding that is often severe enough to require transfusion. The circumstances preceding the event are commonly the imbibing of a large quantity of beer or other alcoholic drink followed by eating and then vomiting. The largest group of individuals affected are alcoholic men. Diagnosis is made by direct vision using an endoscope. Most ruptures spontaneously stop bleeding, and the tear heals over the course of some days. If transfusion does not correct blood loss, surgical suture of the tear may be necessary. An alternative to surgery is the use of the drug vasopressin, which shuts down the blood vessels that supply the mucosa in the region of the tear.

Mallory-Weiss lesion

Cancer. Tumours of the esophagus may be benign or malignant. Generally, benign tumours originate in the submucosal tissues and principally are leiomyomas (tumours composed of smooth muscle tissue) or lipomas (tumours composed of adipose, or fat, tissues). Malignant tumours are either epidermal cancers, made up of unorganized aggregates of cells, or adenocarcinomas, in which there are gland-like formations. Cancers arising from squamous tissues are found at all levels of the organ, whereas adenocarcinomas are more common at the lower end where a number of glands of gastric origin are normally present. Tumours produce difficulty in swallowing, particularly of solid foods; they are much more common in men than in women, and they seem to vary greatly in their worldwide distribution. In North China, for example, the incidence of esophageal cancer in men is 30 times that of white men in the United States and eight times that in black men. The incidence in Europeans is much the same as in the United States, except in France, where it is higher. Some correlations with the alcohol or tobacco consumption of the population have been established, suggesting that continued ingestion or inhalation of these materials may predispose a person to malignant change. In the United Kingdom, for example, there is a higher than average incidence in men working in the alcohol trade. Previous stricture formation also apparently plays a role.

In women, cancer of the upper esophagus is more common than in men, and women may be predisposed by long-standing iron deficiency, or Plummer-Vinson (Paterson-Kelly) syndrome. Dysphagia is the first and most prominent symptom. Later swallowing becomes painful as surrounding structures are involved. Hoarseness indicates that the nerve to the larynx is affected. The diagnosis is suggested by X ray and proved by endoscopy with multiple biopsies from the area of abnormality. Diagnosis can be reinforced by removing quantities of cells with a nylon brush for examination under a microscope (exfoliative cytology). The prognosis is poor because the tumour has usually been growing for one or two years before symptoms are apparent. The channel of the esophagus is encroached upon and can be almost entirely obliterated. The disease is usually accompanied by considerable weight loss, but nutrition may be restored by dilating the lumen through the tumour and passing a flanged tube prosthesis through it so that the proximal end is above the tumour and the distal end is in the stomach. In advanced cases, this may be the only useful measure. Where the channel is greatly narrowed, the tumour can be debulked by destroying the tissue with lasers. Radiotherapy is used for malignancies of the upper esophagus and as treatment for those at the lower end. Some clinics practice a combination of radiotherapy followed by surgical excision. The five-year survival rate remains very low, 15 percent at best. Lessening the effects of the disease, with restoration of eating

Plummer-Vinson syndrome

ability, is very important, because otherwise the inability to swallow even saliva is distressing and the patient dies slowly from starvation.

The poor results of treating cancer of the esophagus have stimulated the development of new techniques based on endoscope-guided laser light beams. The developments include preliminary uptake of a photosensitizing agent by the tumour before its exposure to laser light and the use of a krypton-fluoride laser with ultraviolet wavelength and a power density of 10^8 watts per square centimetre. The extent and depth of the tumour in the wall of the esophagus can be determined using ultrasound probes under endoscopic guidance.

STOMACH

Indigestion. The stomach moves in a rolling and wringing pattern, beginning about one-third of the way down the length of the organ, propelling the mixture of food and juices toward its outlet, the pylorus. Any disorder that affects the power of coordination of the stomach muscles is capable of producing symptoms ranging from those that are mildly unpleasant to others that are life-threatening. The unpleasant sensations, called anorexia and nausea, seem to be mediated through the central nervous system, with reflex input from nerve endings in the stomach and duodenum. Sometimes the entire duration of a nausea-vomiting episode is so short that it appears to be vomiting alone, obscuring the presence of nausea. This is characteristically noted in persons with primary diseases of the brain, especially those with tumours or meningitis in which the cerebrospinal fluid is under increased pressure. In many diseases, vomiting may not be preceded by nausea at all, and in others there may be a long time lag between nausea and vomiting. Seasickness is the best known example of this relationship.

The intrinsic muscles of the stomach are innervated by branches of the vagus nerves, which travel along the esophagus from their point of emergence in the brain stem. Severing these nerves, as is often done in the surgical treatment of peptic ulcer, may produce temporary or more prolonged change in the ability of the stomach to empty itself. Many drugs, particularly the anticholinergic medications, are often used in the treatment of peptic ulcer. These drugs exert an action comparable to that produced by cutting the vagus nerves, but they have the potential disadvantage of reducing the flow of saliva, interfering with vision by disturbing accommodation reflexes in the pupil, and reducing the power of the muscle on which bladder emptying depends.

Treatment of peptic ulcer

Gastric retention may result from the degeneration of the nerves to the stomach that can result from diabetes mellitus. Obstruction due to scarring in the area of the gastric outlet, or to tumours encroaching on the lumen, causes the stomach to fill up with its own secretions as well as with partially digested food. In these circumstances, vomiting leads to dehydration and to electrolyte losses, which threaten life if not corrected. The ingestion of soluble alkali in this situation may aggravate the disturbance in the acid–base balance of the body. Bulimia, a nervous disorder characterized by compulsive eating followed by vomiting and purging, can cause severe dehydration and even a ruptured stomach, and it can prove fatal.

Ulcerative diseases. The area of the stomach in which acid and pepsin are secreted has the highest resistance to peptic ulcer. The mucosa elsewhere is less well protected, and its breakdown may lead to ulceration. If the breach is confined to the superficial layers of the mucosa, it is called an erosion; if it extends through the intrinsic layer of muscle of the mucosa into the tissues below, it is known as an ulcer. Erosions and ulcers can be acute or chronic according to how readily they heal. The circumstances that contribute to mucosal injury and ulcer formation include physical and chemical trauma that result from hot fluids and food, aspirin and other drugs, irritating spices, and pickling fluids.

Reduction of the secretion of mucus by the stomach and duodenum, or a reduction of bicarbonate secretion (which neutralizes excess acid), lowers resistance. Extracellular paracrine messenger substances, in particular the prostaglandins, which stimulate secretion of mucus and bicarbonate and also stimulate cell replication, may be deficient in the mucosa of individuals with peptic ulcer. The rate of cell replication is lowered by malnutrition.

In the United States and the Western world generally, duodenal ulcer is much more common than gastric ulcer, occurs more often in men than in women, and is aggravated by nervous tension and fatigue. In Japan, gastric ulcer is more common than duodenal ulcer and is thought to be related to the raw fish and acetic acid pickles of the traditional diet.

Disturbed physiology

In 60 percent of persons with duodenal ulcer, acid and pepsin secretion is on the high end of normal, and in 40 percent, it is above normal. Of particular concern in duodenal ulcer is the inappropriate secretion of acid and pepsin due to failure of homeostatic-controlling mechanisms, so that it continues when the stomach is empty and during the night hours. In special circumstances such as the state of shock produced by large burns, intracranial surgery, coronary occlusion, and septicemia, acute and rapidly penetrating ulcers may occur.

Genetic factors are involved in the development of ulcers. Inheriting blood group O and an inability to secrete antigen substance H into body fluids, while secreting the Lewis antigen, renders a person four times more likely to develop duodenal ulceration than average. There are families in whom the secretion of pepsinogen I is excessive and renders them prone to duodenal ulcer since excess acid secretion is linked to excess pepsinogen I secretion. Possession of blood group A increases the chance of developing a gastric ulcer. Otherwise, the secretion of acid and pepsin in individuals with gastric ulcer is normal in all respects. There are other distinct differences between ulcers at the two main sites: duodenal ulcer is most common between 25 and 35 years of age, while gastric ulcer is uncommon before 40 years and has a peak frequency between 55 and 65 years.

Pain is the major symptom of duodenal ulcers. The pain is a burning or gnawing sensation felt in the upper abdomen in the midline below the sternum. In gastric ulcer it comes on soon after eating, whereas in duodenal ulcer it comes on when the stomach is empty, one and a half to two hours after meals and during the night hours. In the early stages of the disease, the pain is easily and immediately relieved by alkalis and, in duodenal ulcer, by light food.

Peptic ulcers naturally show remissions and relapse. Attacks last from four to eight weeks before spontaneously ceasing. In most instances this behaviour corresponds to healing and breakdown of the ulcer, but an active ulcer may produce no symptoms. Gastric ulcers are slower than are duodenal ulcers to pass into remission and to respond to treatment. The fluctuating course of activity continues for five to 20 years or more. At some point in this span, the symptoms disappear in 60 percent of the individuals, never to recur. Of the remainder, 20 percent have surgery at some time because of complications and resistance to cure, and 20 percent have some degree of disability throughout their lives.

Response to treatment

Gastric ulcers almost always recur in the same site within the stomach, but duodenal ulcers are often multiple and recurrence may be anywhere in the bulb. Furthermore, duodenal ulcers are usually accompanied by an inflammation affecting the whole of the bulb (duodenitis). Multiple erosions varying in size between 0.5 and five millimetres are frequently scattered over the mucosa. With gastric ulcers the inflammation is usually confined to the immediate vicinity of the crater and, as a rule, is not accompanied by erosions. The exceptions are gastric ulcers in the antrum and prepyloric area associated with the use and abuse of analgesics and nonsteroidal anti-inflammatory drugs for arthritic disorders, in which multiple erosions are commonly present.

The most common site of gastric ulcers is halfway up the inner curvature of the stomach at the junction of the lower one-third with the upper two-thirds. This may be because blood flow to this site is more easily reduced than elsewhere. Chronic gastric ulcers at this site are strongly associated with obstructive disease of the airways (chronic

bronchitis and emphysema). Smoking impairs the healing of both gastric and duodenal ulcers.

Complications

The complications of peptic ulcers are hemorrhage, perforation, and obstruction of the outlet of the stomach (pyloric stenosis) by scarring of the duodenal bulb or of the pyloric channel. Bleeding may be obscured because of oozing from the floor of the ulcer and detectable only by testing the feces with blood-sensitive agents, or bleeding may be brisk, leading to the passage of tar-coloured stools (melena). Occasionally when the ulcer erodes into a large vessel bleeding is torrential and life-threatening. Brisk bleeding is usually accompanied by the vomiting of blood (hematemesis), which requires treatment by blood transfusion. In the elderly, hardening of the arteries (arteriosclerosis) prevents the vessel from closing down around the breach. If bleeding persists or recurs, surgery becomes necessary. Ulcers that penetrate the back wall of the stomach or duodenum erode into the pancreas, and back pain becomes prominent. If the ulcer penetrates the anterior wall, free perforation into the abdominal cavity may occur. This causes immediate, intense pain and shock, and the abdominal wall becomes rigid. In most instances this requires emergency surgery with drainage of the abdomen.

Obstruction of the gastric outlet through the pylorus and duodenum by scarring leads to bouts of vomiting and accompanying malnutrition and requires surgery. The mortality of bleeding is high in the aged because of chronic changes in the lungs, heart, and blood vessels, which reduces cardiorespiratory reserves. This is further aggravated by smoking.

Surgery for chronic ulceration is used less frequently since the introduction of powerful drugs that block the pathways of stimulation of the parietal cells and of others that increase the resistance of the mucosa. Two major classes of drugs are involved. The first class comprises drugs such as cimetidine and ranitidine, H_2-receptor antagonists that engage the histamine receptor sites on the outer membrane of the parietal cells without initiating the production of acid. These drugs are called antagonists because they bind a receptor but do not elicit its normal response. The second class of drugs has no effect on acid secretion but lowers the activity of pepsin by absorption, binds to the tissue in the floor of an ulcer and forms a clot and increases the resistance of the mucosa to break down. This class includes sucralfate and tripotassium dicitrato-bismuthate. A third class, which inhibits the ATPase enzyme inside the cell, preventing the pumping out of the acid ions, is being studied. The most promising agents for treating peptic ulcer diseases are the prostaglandins because they inhibit acid secretion and protect the tissue.

For most individuals, a single tablet of an H_2-receptor antagonist taken before sleeping is adequate to control the disease and can be used for long-term suppression. A proportion of persons taking this drug relapse and need a higher dosage from time to time.

Gastritis. A diffuse inflammation of the stomach lining, gastritis is usually an acute process caused by contaminated food, alcohol abuse, or by bacterial- or viral-induced inflammation of the gastrointestinal tract (gastroenteritis). Such episodes are short-lived and require no specific treatment. Unlike the pain of ulcerative disease, the discomfort is generalized in the upper abdomen and is continuous, but it progressively subsides over two or three days. Aspirin and nonsteroidal anti-inflammatory drugs taken for arthritis cause erosions in the antrum of the stomach, which in some instances becomes chronic ulceration. This usually responds to the withdrawal of the offending drugs and treatment with the same agents used to treat peptic ulcers of the stomach and duodenum. Erosions from drugs are a common cause of bleeding.

Drug erosions

The other form of gastritis is gastric atrophy, in which the thickness of the mucosa is diminished. This may be the culmination of damage to the stomach over many years. Although the mechanism is not understood, there are immunological features. Diffuse gastric atrophy leads to partial loss of the glands and secreting cells throughout the stomach and may be associated with iron-deficiency anemia. Atrophy of the mucosa confined to the body and fundic regions of the stomach is seen in pernicious anemia and is due to the formation of antibodies to intrinsic factor secreted by the parietal cells. Intrinsic factor is necessary to the absorption of vitamin B_{12}.

Cancer. Malignant tumours of the stomach are common, but they show variations in incidence from country to country, probably a result of both genetic and environmental factors. Cancer of the stomach often occurs in older persons whose stomachs produce only small quantities of acid. Whether this indicates that the same process that depresses acid secretion also causes cancer or that gastric cancer is inhibited normally by the secretion of acid and pepsin is not known. Gastric cancer affects men more often than women and accounts for about 20 percent of all deaths from cancers of the gastrointestinal tract in the United States. In Japan, on the other hand, it accounts for nearly 70 percent of such cancers.

Other malignant tumours that involve the stomach are tumours ordinarily made up of lymphoid and connective tissue, respectively. Benign tumours, especially leiomyomas, are common and may, when large, cause massive hemorrhage. Polyps of the stomach are not common except in the presence of gastric atrophy. Treatment for these tumours, benign or malignant, is surgery.

Symptoms of stomach cancer

Because symptoms produced by tumours of the stomach are highly variable, there are no common characteristics of the disease in its early stages. The symptoms most often seen are loss of appetite, some weight loss, and symptoms attributable to an anemia, a condition that frequently is present because of blood loss into the stools, which, though constant, is usually so minimal as to escape notice by the patient. Tumours in the lower part of the stomach produce obstructive symptoms, and tumours high in the stomach may obstruct the esophageal entry into the stomach, producing difficulty in swallowing. Although pain is usually mild, it may be the most noticeable symptom. Stomach cancers often spread to neighbouring lymph nodes or to the liver in the early stages, accounting for the low percentage of cures by surgery.

DUODENUM

The duodenum, aside from being the site of duodenal peptic ulceration, is otherwise not an important seat of disease. It is, however, often involved in the diseases of its neighbours, in particular the pancreas and the biliary tract. Primary cancer of the duodenum is an infrequent disease. Benign tumours, particularly polyps and carcinoids, are more frequent. Cancers of the common bile duct or of the pancreas are important causes of death and may make their presence known by what they do to the duodenum, particularly in terms of obstruction and pain. It is because of their encroachment on the duodenum that these entities often are diagnosed by upper intestinal X-ray studies. Benign anomalies of the organs of this area, like an encircling ring of pancreas, may also encroach upon the duodenum. In countries of the Middle and Far East, where parasites are endemic, roundworms and tapeworms in particular are often found anchored in the duodenum. In inflammations of the pancreas, the neighbouring duodenum is often involved in such a way as to produce impairment of motility and occasionally ulceration with hemorrhage. A protozoal parasite, *Giardia lamblia,* can contaminate drinking water and is a common cause of diarrhea and, if unrecognized, malabsorption.

SMALL INTESTINE

Causes of cramping pain

A lack of coordination of the inner circular and outer longitudinal muscular layers of the intestinal wall usually results in an accumulation of excess contents in the lumen, with consequent distension. This distension may cause pain and usually results in hyperactive contractions of the normal segment next to the distended area. Such contractions may be strenuous enough to produce severe, cramping pain. The most common cause of disturbed motility in the small intestine is food that contains an unsuitable additive, organism, or component.

Traveler's diarrhea. Traveler's diarrhea is watery, accompanied by cramps, and lasts a few days. It is almost always caused by toxin-generating *Escherichia coli,* less often by other organisms. *Shigella* infection may occur

simultaneously, however, and visitors to countries where giardiasis is endemic may suffer infection. Salads remain the most common cause of traveler's diarrhea in countries where the climate is hot. Such diarrhea generally disappears spontaneously with abstention from food accompanied by drinking of nonalcoholic fluids. Mixtures of sodium and potassium chloride, sodium bicarbonate, and glucose reconstituted with water are one method of treatment.

Intestinal obstruction. The most serious problems in small intestine motor disturbances arise from an intestinal obstruction that results from an actual encroachment on the bowel by an adhesive band or from an internal block produced by a tumour or gallstone. As profound an obstruction results when a portion of the intestine undergoes partial necrosis, or death, from failure of its blood supply. This necrotic section cannot pass peristaltic activity and, for all practical purposes, serves as an obstruction. The death of the tissue, furthermore, results in the escape of highly toxic fluids from the intestinal contents through the wall, producing peritonitis. Surgery is usually necessary early in the course of the illness.

Tissue necrosis

The speed with which the contents are passed along the small intestine is increased by many factors, including swallowing air, the nature or physical characteristics of foods (*i.e.*, whether they are cold or highly spiced), and, most importantly, by emotions. An upright position during physical activity causes the small intestine to swing about on its mesenteric root, so that merely assuming the horizontal position often markedly slows intestinal motor activity.

Irritable bowel syndrome. The extremely common disorder known as the irritable bowel syndrome is probably due to a disturbance of the motility of the whole intestinal tract. The symptoms vary from watery diarrhea to constipation and the passage of stools with difficulty. When the colon is involved, an excess of mucus is often observed in the stools. Pain is most often felt in the lower abdomen and in the left lower quadrant. Generalized abdominal discomfort, sometimes with nausea, may follow defecation and may last 15 to 30 minutes. Most sufferers are nervous persons, and some have periods of anxiety depression. Motor activity in those individuals with diarrhea tends to be decreased in the lower colon, whereas motor activity in those with constipation is increased. The strength of contractions is greatly augmented, the frequency increased, and there are periods of diffuse spasm when a long sector of the colon stays contracted.

Occasionally the irritable bowel syndrome may be due to an allergy to a particular foodstuff. The syndrome may develop following an infection such as bacillary dysentery, after which the small intestine remains irritable for many months. Treatment of the irritable bowel syndrome is best limited to an explanation to the patient, elimination of stress, psychological support, change in life-style if appropriate, and outdoor activities. Possible aggravating items such as lactose-containing foods, coffee, and deep-fried dishes should be eliminated from the diet, and dietary fibre should be added to help in modulating motility and resolving the diarrhea and the costive activity of the condition. When discomfort is prominent, antispasmodic agents that relax smooth muscle, such as dicyclomine hydrochloride or mebeverine, may be prescribed. If diarrhea does not respond to an increase in dietary fibre, diphenoxylate or loperamide may slow the movement of the intestinal contents, thereby increasing the potential for the reabsorption of water.

Malabsorption. Of the mixture of fats, proteins, carbohydrates, minerals, and vitamins that are ingested, it appears that fat is most subject to malabsorption. Measurements of fat absorption have been used for some time as an index of general intestinal malabsorption. Fat must be solubilized to be absorbed, and this requires optimum conditions for the action of bile salts, phospholipids, and lipase. Malabsorption occurs when the small intestine is unable to transport properly broken down products of digestive materials from the lumen of the intestine into the lymphatics or mesenteric veins, where they are distributed to the rest of the body. Defects in transport occur either because the absorptive cells of the intestine lack certain enzymes, whether by birth defect or by acquired disease, or because they are hindered in their work by other disease processes that infiltrate the tissues, disturb motility, permit bacteria to overpopulate the bowel, or block the pathways over which transport normally proceeds.

The basis of malabsorption

Diagnosis of malabsorption is made primarily from the patient's history, physical examination, X-ray films of the abdomen, and study of the stools under controlled dietary conditions. Some test substances can be administered and their recovery measured in the stools and in the blood. Motor aspects of the intestine can be studied using a variety of techniques.

Congenital malformations. Meckel's diverticulum is a common congenital malformation that occurs when the duct leading from the navel to the small intestine in the fetus fails to atrophy and close. The duct serves the fetus as the principal channel for nourishment from the mother. The diverticulum in the child or adult may range from a small opening to a tube that is a foot or more in length and it may contain cells derived from the stomach glands that secrete acid and pepsin. If such secretions spill onto normal intestinal mucosa, which is totally nonresistant to it, the mucosa ulcerates and often bleeds. Thus a peptic ulcer can develop at a site far distant from the stomach or duodenum. The peptic ulcer gives rise to pain, bleeding, or obstruction, and it is the most common cause of bleeding from the lower intestine in children. Meckel's diverticulum must be treated surgically.

Meckel's diverticulum

Another congenital problem in the small intestine is the presence of multiple diverticula, or outpouchings of mucosa and serosa. These are seen usually in elderly persons, although occasionally one may be the site of acute inflammation in a young adult. Bacteria flourish in these diverticula because the outpouchings have no motor activity and cannot empty themselves. The bacteria deprive the body of nutrients and may cause diarrhea and serious malabsorption. The overgrowth of bacteria also upsets the motor activity of the small intestine. Antibiotics may control the condition in the elderly, but surgical resection of diverticula is used in younger persons.

A third congenital malformation is a failure of complete rotation of the small and large intestine, which is a normal step in the development of the fetus. This can result in abnormal intestinal attachments with a subsequent risk of obstruction when the intestine twists around the attachments.

Bacterial infections. Besides the *Shigella* infections described above, other organisms can infest the human body and cause disease. Species of *Salmonella* that cause typhoid and paratyphoid remain endemic scourges in the hot countries and, together with *Shigella*, are occasional causes of epidemics in institutions, especially where the elderly are concerned. The diagnosis is confirmed by the presence of the organisms after a stool culture. Antibiotics and solutions rich in electrolytes are effective therapy. Only severe cases of *Shigella* infection require antibiotics, but typhoid and paratyphoid require hospitalization and, if severe, treatment with chloramphenicol, sulfonamides, or antibiotics. Periodic injection of vaccine is advisable for the protection of individuals exposed to areas where typhoid and paratyphoid are endemic.

Cholera, caused by *Vibrio cholerae*, is endemic to Southeast Asia and periodically becomes pandemic (widely distributed in more than one country). The oral or intravenous administration of electrolyte solutions rich in potassium has revolutionized the treatment of cholera, because deaths are due to a massive depletion of electrolytes and water. The toxin produced by *V. cholerae* attaches to the intestinal cells, the enterocytes, where it stimulates the membrane enzyme adenylate cyclase; this in turn interferes with the intracellular enzyme 3′,5′-cyclic adenosine monophosphate synthetase (cyclic AMP), thereby disrupting the sodium pump system for movement of water and allowing potassium and bicarbonate to seep out of the cell.

Cholera

Parasitic infections. In hot countries, parasitism is endemic. Roundworms, tapeworms, amoebae, hookworms, strongyloides, threadworms, and blood flukes (schistosomiasis) are the main types of parasites. Consequently it is commonplace in these areas for multiple parasite in-

festation to occur in addition to other disorders. This phenomenon, reflecting poverty, lack of education, malnutrition, contaminated drinking water, and inadequate sanitation, is a major factor in chronic ill health and early death.

Parasitic infections. *Roundworms.* Roundworms, or *Ascariasis lumbricoides,* may cause intestinal obstruction if present in sufficient numbers. As they mature from the larval state to the adult worm, they migrate through the body, causing fevers, pneumonitis (lung inflammation), cholangitis (inflammation of the bile ducts), and pancreatitis. Roundworms interfere with the absorption of fat and protein in the intestine, which causes diarrhea. They are eliminated with the administration of piperazine or other anthelmintics, but occasionally surgery is required for obstruction.

Hookworms. Hookworm, or *Ancylostoma duodenale,* infection begins when the worm is in the larval stage. It penetrates the skin, usually of the feet, and migrates during its life cycle through the liver and the lungs, and it attaches to the mucosa of the small intestine where it matures. Hookworms deplete the body of nutrients, and a major effect is severe chronic iron-deficiency anemia. This effect can be corrected with the oral administration of iron, and the number of worms can be controlled with tetrachloroethylene or other anthelmintics.

Life cycle of the hookworm

Threadworms. Threadworms, or *Enterobius vermicularis,* live mainly in the cecum. The adult female migrates at night to the anus and lays eggs on the perianal skin, which causes anal itching. Transmission of the threadworm is fecal–oral, and it can affect an entire family. Threadworms can be eradicated with piperazine or vyprinium embonate.

Tapeworms. The common tapeworms are *Taenia saginata,* found in beef, and *T. solium,* found in pork. Larvae of *Echinococcus granulosus, Diphyllobothrium* species, and some dwarf tapeworms also cause disease. Fertilized ova are passed in feces and are ingested by an intermediary host animal, such as a cow. The embryos migrate to the bloodstream and on reaching muscle or viscera develop into larvae. When the flesh is consumed by humans, the larvae pass into the intestine, where they attach and mature into adult worms. Thus the most common source of infection is inadequately cooked meat. Tapeworms found in beef and pork only give rise to symptoms if their number and size cause intestinal obstruction. *Diphyllobothrium latum,* a fish tapeworm, may cause a severe anemia similar to pernicious anemia, because it consumes most of the vitamin B_{12} in the diet of the host.

Appendicitis. Appendicitis is an inflammation of the vermiform appendix that may be caused by infection or partial or total obstruction. It is a major cause of intra-abdominal pain, principally attacking those younger than 35 years of age. Appendicitis is easily diagnosed and is treated with surgery. Widespread use of antibiotics for upper respiratory and other diseases may have lessened the incidence of the acute form of this disorder, so that more cases of late-developing appendiceal abscess are being reported. Parasitic worms also can contribute to its incidence. Appendicitis occasionally occurs in elderly people, and instances where an abscess forms and bursts require urgent surgery.

Chronic inflammations. Chronic inflammations of the small intestine include tuberculosis and regional enteritis (Crohn's disease). These disturbances are difficult to diagnose in their early stages since their initial symptoms are often vague. General symptoms include low-grade fever, a tendency toward loose stools, weight loss, and episodes of cramping abdominal pain caused by obstruction of the lumen and interference with normal muscular activity by the inflammation of the intestinal wall. Diagnosis is usually made by X ray. There is specific drug therapy for tuberculosis. In Crohn's disease anti-inflammatory, nonspecific drugs are helpful. Surgical excision of the diseased segments of intestine may be necessary.

Crohn's disease

The incidence of Crohn's disease is rising. The disease is largely confined to young women and may reflect the increase in smoking and use of oral contraceptives. About 60 percent of persons with Crohn's disease require surgery because of obstruction of the intestinal lumen and another 20 percent because of fistulation, or connection, between adjacent structures, for example, from the sigmoid colon to the bladder. A combination of repeated surgical excisions from the small intestine and disease of the intestinal wall can result in a severe malabsorptive state. This sometimes requires long-term intravenous (parenteral) nutrition. Crohn's disease tends to affect the colon more often now than previously.

Celiac disease. Celiac disease affects between one in 500 and one in 2,000 persons, depending on the region of the world. The disease is due to an inability to tolerate a fraction of the gluten, a protein present in wheat, rye, barley, and some oats. It is not clear whether the disability is an allergy, with the damage to the mucosa of the small intestine caused by an immune reaction, or whether it is caused by an inability to break down a toxic protein, gluten, to smaller peptide fractions. The existence of a specific enzyme deficiency has not been proved.

Cause of celiac disease

Studies of the immune function of those with celiac disease suggest that at least a major part of the process is a delayed hypersensitivity reaction and that the morphological changes are correlated with the presence of circulating antibodies to gluten. The mucosal reaction results in progressive atrophy, with dwarfing, if not complete disappearance, of the microvilli and villi that line the intestinal tract. This dramatically reduces the area available for absorption, and malabsorptive diarrhea results. The peak frequency is between six and 24 months of age but the disorder may not manifest itself until middle age, or, if mild, may be unnoticed until then. Iron- and folic-acid deficiency anemias, softening of the bones (osteomalacia), and general weakness may be accompanied by a variety of disorders attributable to the nonabsorption of vitamins. Untreated, it is a serious though rarely life-threatening disease after infancy. Diagnosis is established by a biopsy. Withdrawal from the diet of the cereals that contain gluten generally brings about dramatic improvement and disappearance of all symptoms. A few cases require treatment by corticosteroids as well. Untreated celiac disease may be complicated in middle age by malignant lymphoma of the jejunum. A skin condition, dermatitis herpetiformis, seems likely to be a variant of the disease, and some patients respond immediately to a gluten-free regimen.

Tropical sprue. A malabsorption disorder of unknown cause, tropical sprue affects residents and visitors in tropical countries. It is associated with partial atrophy of the mucosa of the small intestine. Its symptoms are diarrhea, anorexia, and fatigue. If the disease is prolonged, anemia caused by malabsorption of vitamin B_{12} develops. Steatorrhea (excess fat in stools) is common, and glucose absorption is impaired. Prolonged treatment with antibiotics, such as tetracycline, and the replacement of vitamins, especially B_{12} and folic acid, are successful.

LARGE INTESTINE

A wide variety of diseases and disorders occur in the large intestine. Imperfect fetal development may result in an anus that has no opening, a defect that requires major plastic surgery to correct. Abnormal rotation of the colon is fairly frequent and occasionally leads to disorders. Unusually long mesenteries (the supporting tissues of the large intestine) may permit recurrent twisting, cutting off the blood supply to the involved loop. The loop itself may be completely obstructed by rotation. Such complications are usually seen in elderly patients and particularly in those with a long history of constipation.

Simple constipation. Brain disease, metabolic failure, or drugs can dull the normal signals that give rise to the urge to defecate. Poor abdominal musculature or a poor pelvic floor, sometimes the result of surgery or childbirth, makes it difficult to mobilize effective pressures to bring about defecation.

Congenital megacolon. A disease that is analogous to achalasia of the esophagus is an idiopathic (the cause being unknown) condition called aganglionic megacolon, or Hirschsprung's disease. It is characterized by the absence of ganglion cells and normal nerve fibres from the distal (or lower) three to 40 centimetres of the large intestine.

Aganglionic megacolon

Neuromuscular transmission is absent from this segment, and peristalsis cannot occur. It is thus a functional obstruction. In 10 percent of cases a larger segment is involved and, on rare occasions, the whole colon. The area of normal intestine above the obstruction works harder to push on the fecal contents, and eventually the muscle of the normal segment thickens. The entire colon thus slowly becomes more and more distended and thick-walled. Diagnosis is made by the microscopic appearances in a deep biopsy of the lower rectum. Various surgical procedures are used to correct the condition.

Acquired megacolon. Acquired megacolon is commonly caused by a combination of faulty toilet training and emotional disorders during childhood, in which the child withholds defecation. This starts a cycle of the administration of increasing amounts of laxatives with, ultimately, damage to the intrinsic innervation in the intestinal wall. A huge, dilated rectum full of feces develops over the years. The impacted feces act as an obstruction, and further fecal material piles up behind, with voluminous dilatation of the whole colon in some cases. The task of the evacuation of the contents of the bowel prior to surgery, if it is required, may require hospitalization for up to three months. The same phenomenon is occasionally encountered in those with schizophrenia and severe depression. It may be related to neurological disorders such as paraplegia, to unrecognized rectal strictures, and to some metabolic disorders. Severe degrees of constipation, often running in families and leading to megacolon, occur, but the cause has not been discovered. Persons with the disorder are often mistakenly regarded as psychoneurotic or malingerers. Resection of the colon and uniting the ileum to the rectum is effective treatment.

Diarrhea. Because water is normally absorbed from the colonic content, principally in the ascending, or right, colon, any inflammatory, neoplastic, or vascular disturbance of that part of the colon usually decreases the firmness of the stool, increases its 24-hour total weight, and produces blood or other evidence of inflammation.

Lactase deficiency

An example of the relation of motor and absorptive defects is shown by those who are deficient in lactase, the enzyme that splits lactose (milk sugar) into its component parts, glucose and galactose. Shortly after drinking milk, such persons usually have severe intestinal cramping, followed later by watery diarrhea. The lactose in the milk is not broken down, and it stays in the lumen of the small intestine, drawing water to it. The increased bulk of fluid and sugar distends the intestine, which then contracts actively. The rapid contractions become crampy, and the material is driven along the intestine into the colon, which cannot absorb the water rapidly enough. The resultant watery, unformed stools are frequently acidic.

Gas. Gas–fluid mixtures do not provide the distribution of peristaltic pressure as smoothly as do fluids alone. When a person is in an upright position, gas diffuses to the uppermost portions of the colon: the right and left "corners." There it is compressed by the contraction of adjacent segments, giving rise to pain that is localized either near the liver and gallbladder or under the diaphragm and heart. This pain can be incorrectly thought to be associated with diseases of these organs whereas it is actually caused by increased gas in the colon. Eating slower to reduce the amount of air ingested, decreasing the intake of carbonated beverages and whipped desserts that contain air bubbles, and avoiding certain gas-producing foods, such as most beans, onions, sprouts, nuts, and raisins, usually help to reduce flatulence.

Diverticula. Arteries penetrate the muscular walls of the colon from its outside covering, the serosa, and distribute themselves in the submucosa. The channels in which these arteries lie may be regarded as potential tunnels for hernias. With aging, and perhaps in persons predisposed to the disorder, these channels become larger. If the peristaltic activity of the colon maintains a high pressure within its lumen, as in patients straining to defecate, the mucous membrane of the colon may be driven slowly into these channels and eventually may follow the arteries back to their site of colonic entrance in the serosa. At such time, the outward-pushing mucosa becomes a budding sac, or diverticulum, on the antimesenteric border of the colon but with a connection to the lumen. In the Western world, multiple colonic diverticula occur in as many as 30 percent of persons older than 50 years. Diverticula are particularly common in those whose diets are deficient in fibre (roughage), and they are rare in countries where high-fibre diets are usual. Hypertrophy (increase in size and mass) of the muscle fibre of the colon, especially in the sigmoid region, precedes or accompanies diverticulosis; this is especially apparent in the diverticulosis in middle-aged persons as opposed to that in the elderly.

Diverticulosis

The principal dangers of diverticulosis are massive hemorrhage and inflammation. Hemorrhage results from the action of hard stools against the small arteries of the colon that are exposed and unsupported because of diverticula. As the arteries age, they become less elastic, less able to contract after bleeding begins, and more susceptible to damage. Diverticulitis, on the other hand, occurs when the narrow necks of the diverticula become plugged with debris or inedible foodstuff and bacteria, uninhibited by the usual motor activity that keeps the intestine clean, proliferate in the blind sacs. When the sacs enlarge, the adjacent intestinal wall becomes inflamed and irritable, muscle spasms occur, and the patient experiences abdominal pain and fever. If the sacs continue to enlarge, they may rupture into the peritoneum, giving rise to peritonitis, or an inflammation of the peritoneum. More commonly they fix themselves to neighbouring organs and produce localized abscesses, which may prove difficult to treat surgically. Mild diverticulitis responds well to conservative treatment and to antibiotics; massive hemorrhages often require emergency surgery. Recurrent diverticulitis requires resection of the affected area of the colon.

Abscesses. Abscesses (cavities of pus formed from disintegrating tissue) in the perianal area are common complicating features of many diseases and disorders of the large intestine. Fungal infections of the moist and poorly cleansed area around the anus are common and permit the maceration (or gradual breaking down) of tissue and the invasion by bacteria from the skin and colon. In a diabetic, who is susceptible to skin infection, scrupulous perianal hygiene is very important.

Bacterial infections. The colon may become inflamed because of invasion by pathogenic, or disease-causing, bacteria or parasites. A variety of species of *Shigella,* for example, attack the mucous membrane of the colon and produce an intense but rather superficial hemorrhage. In infants and in the elderly, the amount of fluid and protein lost by the intense inflammatory response may be fatal, but ordinarily such symptoms are less serious in otherwise healthy persons. *Salmonella* species, responsible for severe generalized infections originating from invasion of the small intestine, may damage the lymph follicles of the colon, but they do not produce a generalized inflammation of the colon (colitis). The cytomegalic virus can cause a severe colitis, producing ulcerations. Lymphopathia venereum causes a more generalized and superficial colitis.

Food residues provide an excellent culture medium for bacteria, and the interior of the colon is a nearly ideal environment for their growth. The most important parasite producing disease in the human colon is the protozoan *Entamoeba histolytica.* This widely distributed parasite enters the human digestive tract via the mouth and lodges in the cecum and ascending colon. This usually results in irritability of the right colon and failure to absorb water properly, so that intermittent, watery diarrhea ensues. The amoebas undermine the mucosal coat and may create large ulcerations that bleed impressively. Stools contain blood, but there is little pus or other evidence of reaction by the colon to the invading organism. In more generalized amoebic colitis, the rectum and sigmoid colon are invaded by *E. histolytica,* which manifest their presence by numerous discrete ulcerations separated from each other by a relatively normal-appearing mucous membrane. The amoebas may enter the portal circulation and be carried to the liver, where abscesses form and sometimes rupture into the chest or the abdominal cavity. Immunologic tests of the blood may help in diagnosis. After identification

Amoebic disease

of the parasites by direct smears from the margin of the ulcers, or from the stools, treatment is with a combination of amoebicidal drugs and a broad-spectrum antibiotic—*i.e.*, an antibiotic that is toxic to a wide variety of parasites—usually metronidazole and tetracycline.

Colitis. The most common form of chronic colitis (inflammation of the colon) in the Western world, ulcerative colitis, is idiopathic (*i.e.*, of unknown cause). It varies from a mild inflammation of the mucosa of the rectum, giving rise to excessive mucus and some spotting of blood in the stools, to a severe, sudden, intense illness, with destruction of a large part of the colonic mucosa, considerable blood loss, toxemia and, less commonly, perforation. The most common variety affects only the rectum and sigmoid colon and is characterized by diarrhea and the passage of mucus. The disorder tends to follow a remitting-relapsing course. About 15 percent of cases of all colitis involve extension of the disease beyond the area initially affected, with an increase in severity. Where the destruction has been extensive, there is a risk of malignancy 10 to 20 years after the onset of the disease.

Crohn's disease. The cause of Crohn's disease is unknown. Apart from the greater tendency for fistulas to form and for the wall of the intestine to thicken until the channel is obstructed, it is distinguishable from ulcerative colitis by microscopic findings. In Crohn's disease, the maximum damage occurs beneath the mucosa, and lymphoid conglomerations, known as granulomata, are formed in the submucosa. Crohn's disease attacks the perianal tissues more often than does ulcerative colitis. Although these two diseases are not common, they are disabling.

Sites of damage

Because there is no specific etiology, a combination of anti-inflammatory drugs, including corticosteroids and aminosalicylic acid compounds, is used to treat Crohn's disease. The drugs are effective both in treating acute episodes and in suppressing the disease over the long term. Depending on the circumstances, hematinics, vitamins, high-protein diets, and blood transfusions are also used. Surgical resection of the portion of the large bowel affected is often done. The entire colon may have to be removed and the small intestine brought out to the skin as an ileostomy, an opening to serve as a substitute for the anus. In ulcerative colitis, as opposed to Crohn's disease, the rectal muscle may be preserved and the ileum brought through it and joined to the anus.

Tumours of the colon are usually polyps (growths from the mucous membranes) or cancers. The tendency of some persons to form polyps is strikingly exemplified in the rare disorder known as familial polyposis, in which the colon may be studded with hundreds or thousands of small polyps. Because a colon that produces so many polyps eventually produces cancers as well, these colons should be removed surgically as soon as the diagnosis is made. The rectum may be left, but a visual examination of the residual mucosa must be made twice yearly to detect signs of early cancerous change. Another peculiar form of polyp is the villous adenoma, often a slowly growing, fernlike structure that spreads along the surface of the colon for some distance. It can recur after being locally resected, or it can develop into a cancer.

Cancer. In the West, cancer of the colon is a more common tumour than is cancer of the stomach, and it occurs about equally in both sexes. Symptoms are highly variable, the main feature being blood in the stools, but this may be detectable only by chemical testing. Cancers compress the colonic lumen to produce obstruction, they attach to neighbouring structures to produce pain, and they perforate to give rise to peritonitis. Cancers also may metastasize to distant organs before local symptoms appear. Nevertheless, the prognosis for patients with this tumour is considerably better than it is for cancer of the stomach. About half the patients who have a colonic cancer removed surgically live at least five years. A colostomy is required for some patients, in which an opening is made from the colon to the skin, where the fecal contents are extruded. After the colon has been removed partially, it is possible to join the terminal ileum or the remnant colon directly to the anal canal. A reservoir also can be fashioned out of the terminal ileum and placed inside the rectum muscle from which the inflamed mucosa has been removed. This functions as a normal rectum and with retained sphincters at the anus can render the patient continent, although there usually are three or four movements daily.

Symptoms of colon cancer

Anal disorders. Anorectal disorders related to defecation are more common in the Western world than elsewhere. Whether this distribution is related to diet, exercise, personal hygiene, or to social customs that inhibit the natural gratification of the urge to defecate is not clear. These disorders usually take the form of fissures (cuts or cracks in the skin or mucous membrane) at the junction of the anal mucous membrane with the skin between the thighs. If such fissures become chronically infected and resistant to treatment by sitz baths and local medication, they may require surgical correction. Anal fistulas sometimes occur as complications of serious bowel disease, as in tuberculosis or Crohn's disease of the bowel, or in certain parasitic diseases. A more general disorder is the enlargement of veins of the rectum and anus to form external or internal hemorrhoids. Many adults in the West have such venous enlargements, but only a small number suffer serious symptoms from their presence. Hemorrhoids protrude, are associated with anal itching and pain, and bleed, especially when they come in contact with hard stools. These symptoms generally can be controlled by conservative measures, but occasionally they persist or cause so much distress that surgical removal of the enlarged and dilated veins is necessary.

Hemorrhoids

In the past, poor obstetrical and postnatal care resulted in a severe loss of support of the pelvic floor in parturient women; prolapse, or downward displacement, of the rectum, as well as of the uterus and bladder, was common. Such complications have largely been eliminated except in conditions such as fibrocystic disease of the pancreas in children or in certain neurological disorders in aged persons.

(W.S.)

LIVER

A variety of agents, including viruses, drugs, environmental pollutants, genetic disorders, and systemic diseases, can affect the liver. The resulting disorders usually affect one of the three functional components of the liver: the hepatocyte (liver cell) itself, the bile secretory (cholangiolar) apparatus, or the blood vascular system. Although an agent tends to cause initial damage in only one of these areas, the resulting disease may in time also involve other components. Thus, although viral hepatitis (inflammation of the liver) predominantly affects hepatocytes, it commonly leads eventually to canalicular damage.

Most acute liver diseases are self-limited, and liver functioning returns to normal once the causes are removed or eliminated. In some cases, however, the acute disease process destroys massive areas of liver tissue in a short time, leading to extensive death (necrosis) of hepatic cells and often to death of the patient. Hepatitis may result from viral infections or toxic damage from drugs or poisons. When acute hepatitis lasts for six months or more, a slow but progressive destruction of the surrounding liver cells and bile ducts occurs, a stage called chronic active hepatitis. If hepatocellular damage is severe enough to destroy entire acini (clusters of lobules), they are often replaced with fibrous scar tissue. Bile canaliculi and hepatocytes regenerate in an irregular fashion adjacent to the scar tissue and result in a chronic condition called cirrhosis of the liver. Where inflammatory activity continues after the onset of cirrhosis, the disorderly regeneration of hepatocytes and cholangioles may lead to the development of hepatocellular or cholangiolar cancer.

Chronic active hepatitis

Acute hepatocellular hepatitis. Although a number of viruses affect the liver, including the cytomegalovirus of infancy and childhood and the Epstein-Barr virus of infectious mononucleosis, there are three distinctive transmissible viruses that are specifically known to cause acute damage to liver cells: hepatitis virus A (HAV), hepatitis virus B (HBV), and hepatitis virus non-A, non-B (NANB).

The hepatitis A virus is transmitted almost exclusively by the fecal–oral route, and it thrives in areas where sanitation

and food handling are poor and hand washing is infrequent. Hepatitis A virus proliferates in the intestinal tract during the two weeks following the onset of symptoms, but it then disappears. Many infected persons are unaware of being ill, since their disease remains asymptomatic or quite mild. The incubation period of HAV infections, from viral ingestion to the onset of symptoms, averages four to five weeks. Acute illness in an otherwise healthy pregnant woman does not appear to have adverse effects upon the fetus. Persons can become passively immunized against hepatitis A attacks for several months with a single injection of immunoglobulin, a product made from pools of 100 or more donor plasmas. Persons can be actively immunized to HAV by acquiring the virus subsequent to becoming passively immunized, but such infections are either inapparent or very mild. An active vaccine is not available, and there are no carriers of the virus.

Incubation period of HAV infections

Hepatitis B virus is present throughout the world in asymptomatic human carriers who may or may not have ongoing liver disease. Formerly, the disease was widely spread by the transfusion of whole blood or blood products, such as the cryoprecipitate used in the treatment of hemophilia. Since the markers of infection have become so readily identifiable, this mode of transmission is much less common, comprising only about 10 percent of cases, compared with 60 percent in the past. Hepatitis B virus is still transmitted in some blood specimens because the levels of virus particles present may be too low to be detected. Virus particles in carriers are found in bodily secretions, especially saliva and sexual emissions, as well as in blood. The incidence of B antigens is high among persons engaging in promiscuous sexual activity, drug addicts who share syringes, health care workers, and infants of mothers who are carriers. Many newly infected persons develop the acute disease within three weeks to six months after exposure, while some develop an asymptomatic form of hepatitis that may appear only as chronic disease years later. Others eliminate the virus completely without any symptoms beyond the appearance of antibodies to surface antigen, while still others become carriers of surface antigen and thus presumably are infective to others.

There are two methods of preventing hepatitis B: passive immunization, through the use of a specific immunoglobulin derived from patients who have successfully overcome an acute HBV infection; and active immunization, through the injection of noninfective, purified HBV surface antigen. The first method is used following specific exposures that carry a high risk of infection, such as using needles contaminated with HBV particles, the ingestion of body products likely to be infected, or the birth of an infant to a surface-antigen-positive mother. The second method, active immunization, is used for those who belong to groups with a high risk of HBV infection, such as children living in endemic areas, medical personnel in high-risk specialties, drug addicts, sexually promiscuous persons, and family groups living close to known carriers. Active immunization, involving a series of three injections of vaccine over a period of three to six months, has been shown to confer a high degree of resistance to infection.

Groups at high risk

Non-A, non-B hepatitis virus has not been isolated, so that the markers of infection and of immunity are not available. Because of this, NANB is the major cause of posttransfusion hepatitis, and it appears with a frequency of three to six cases per 1,000 transfusions of blood prescreened for HBV. The average incubation period of the disease is about seven weeks, and an acute attack of NANB hepatitis is usually less severe than acute hepatitis B. Non-A, non-B hepatitis, however, is more likely to become chronic than is hepatitis B, and it may recur episodically with acute flares.

The symptoms characteristic of the acute hepatitis caused by the HAV, HBV, and NANB viruses are essentially indistinguishable from one another. Patients often complain of a flulike illness for several days, with chills, variable degrees of fever, headache, cough, nausea, occasional diarrhea, and pronounced malaise. Abdominal pain caused by swelling of the liver is a common complaint. As many as half of the infected patients develop only mild symptoms or none at all. A small percentage of patients, especially those with HBV infections, may develop hives, painful skin nodules, acute arthritis, or urinary bleeding caused by the deposition of large immune antigen–antibody complexes in the small blood vessels of adjacent organs. After several days or a week of such symptoms, jaundice commonly develops. At times the jaundice is so mild that it is not noticed by patients, although they often do note that the urine has become dark amber in colour because of the high levels of water-soluble bilirubin transmitted to the kidneys by the bloodstream.

The onset of jaundice usually brings with it a marked improvement in other symptoms (see below, *Biliary tract: Jaundice*). Jaundice lasts about two weeks but may continue for several months, even in those who have complete recovery. Some patients complain of itching during this period, and they notice the light colour of their stools. These symptoms probably result from the compression of bile canaliculi and intralobular bile ducts by the swelling of hepatocytes and Kupffer cells. The changes result in the reduced secretion of bile pigments into the biliary system, their reflux into the bloodstream, and the deposition of bile salts and other biliary constituents in the skin and subcutaneous tissues, a condition called obstructive jaundice. After the phase of jaundice subsides, almost all patients with hepatitis A, and at least 90 percent of those with hepatitis B, recover completely.

The onset of jaundice

Aside from jaundice, the physical examination of patients with acute viral hepatitis may reveal nothing more than the swelling of lymph nodes in the neck. Many patients have detectable enlargement and, at times, tenderness of the liver. Some also show an enlarged spleen. Signs of confusion or disorientation indicate severe damage to the liver. The diagnosis of hepatitis is confirmed by blood tests that show marked elevations of enzymes (aminotransferases) released from damaged liver cells and, at times, by the presence of viral antigens or acute viral antibodies (IgM).

A small number, perhaps 1 percent, of patients with viral hepatitis, especially the elderly, develop a sudden, severe (fulminant) form of hepatic necrosis that can lead to death. In this form of the disease jaundice increases to high levels during the first seven to 10 days, spontaneous bleeding occurs because of reductions of blood-clotting proteins, and irrational behaviour, confusion, or coma follow, caused by the accumulation in the central nervous system of the breakdown products of protein normally metabolized by the liver. Beyond supportive measures there is no effective treatment of fulminant hepatic failure.

Acute hepatitis also may be caused by the overconsumption of alcohol or other poisons, such as commercial solvents (*e.g.*, carbon tetrachloride), acetaminophen, and certain fungi. Such agents are believed to cause hepatitis when the formation of their toxic intermediate metabolites in the liver cell (phase I reactions) is beyond the capacity of the hepatocyte to conjugate, or join them with another substance for detoxification (phase II reactions) and excretion. As long as the levels of these agents are small enough to permit complete phase I and phase II reactions, there is no damage to the liver cell.

Other causes of acute hepatitis

Acute canalicular (cholestatic) hepatitis. This form of hepatitis is most commonly caused by certain drugs, such as chlorpromazine, that lead to idiosyncratic reactions or, at times, by hepatitis viruses. The symptoms are generally those of biliary obstruction and include itching, jaundice, and light-coloured stools. Drug-induced cholestasis almost invariably disappears within days or weeks after exposure to the agent is discontinued. Acute congestive liver disease usually results from the sudden engorgement of the liver by fluids after congestive heart failure. The liver may enlarge and become tender. The levels of hepatocytic enzymes in the blood are often greatly increased, and recovery is rapid once the heart failure improves. Jaundice is uncommon in acute hepatic congestion.

Chronic active hepatitis. The result of unresolved acute injury, chronic active hepatitis is associated with ongoing liver damage. A milder form of chronic disease, called persistent hepatitis, does not appear to lead to progressive liver damage despite evidence of a continuing mild inflammation. These conditions may result from viral hepatitis,

drug-induced hepatitis, autoimmune liver diseases (lupoid hepatitis), or congenital abnormalities. A prominent autoimmune liver disease is Wilson's disease, which is caused by abnormal deposits of large amounts of copper in the liver. Granulomatous hepatitis, a condition in which localized areas of inflammation (granulomas) appear in any portion of the liver lobule, is a type of inflammatory disorder associated with many systemic diseases, including tuberculosis, sarcoidosis, schistosomiasis, and certain drug reactions. Granulomatous hepatitis rarely leads to serious interference with hepatic function, although it is often chronic.

Chronic active viral hepatitis cannot be treated effectively and the course of the disease is usually slow but relentlessly progressive. Cirrhosis of the liver, and occasionally hepatocellular cancer, usually result from a gradual loss of liver function. Chronic active hepatitis that is the result of autoimmune disorders usually responds to the administration of adrenal corticosteroids, which moderate the inflammatory reaction.

Cirrhosis. The end result of many forms of chronic liver injury is cirrhosis, or scarring of liver tissue in reponse to previous acinar necrosis and irregular regeneration of liver nodules and bile ducts. Among the congenital disorders producing cirrhosis are Wilson's disease, hemochromatosis (over-deposition of iron pigment), cystic fibrosis, biliary atresia (congenital absence of a part of the bile ducts), and alpha$_1$-antitrypsin deficiency, or the congenital absence of a proteolytic enzyme inhibitor that results in the accumulation of abnormal forms of carbohydrate in hepatocytes. In the West, cirrhosis of the liver most commonly results from chronic heavy intake of alcohol. Chronic viral hepatitis is probably the leading cause of cirrhosis in underdeveloped countries. Primary biliary cirrhosis, a widespread, though uncommon, autoimmune inflammatory disease of bile ducts, is a disorder primarily affecting middle-aged and older women. The inflammation leads to necrosis and gradual disappearance of bile ducts over a period of one or more decades. Secondary biliary cirrhosis results from chronic obstruction or recurrent infection in the extrahepatic bile ducts caused by strictures, gallstones, or tumours. Infestation of the biliary tract with a liver fluke, *Clonorchis sinensis*, is a cause of secondary biliary cirrhosis in Asia. Cirrhosis occasionally is the result of chronic vascular congestion of the liver in persons with prolonged heart failure and in those with chronic obstruction of the hepatic veins caused by benign blood clots or metastatic cancer. Symptoms of cirrhosis are usually absent during the early stages of the disease. Occasionally, cirrhosis is detected during a physical examination when an enlargement of the liver, spleen, or veins in the upper abdominal wall is found. More often, patients develop symptoms related either to the failure of the liver to perform its functions or to complications caused by the circulatory changes that a cirrhotic liver imposes on the venous blood flow from the intestinal tract (portal hypertension). Thus, common symptoms include jaundice, resulting from reduced passage of conjugated bilirubin into the biliary tract; increased bleeding, from sequestration of blood platelets in a congested spleen; or the deficient production of short-lived coagulation proteins by the liver. In males, there may be certain changes in the skin, such as the appearance of small spider-like vascular lesions on the hands, arms, or face, a marked reddening of portions of the palms, or enlargement of the breast or reduction in testicular size. These symptoms are believed to occur because of the liver's inability to metabolize the female sex hormones normally produced by the body. The gradual accumulation of fluid in the abdominal cavity (ascites), sometimes accompanied by swelling of the ankles, is attributable to portal hypertension and to reduced hepatic production of albumin, while failure of the liver to metabolize amino acids and other products of protein digestion may lead to the state of confusion called hepatic encephalopathy. Loss of appetite, reduction of muscle mass, nausea, vomiting, abdominal pain, and weakness are other symptoms of hepatic cirrhosis. Diabetes in a patient with cirrhosis is caused by hemochromatosis (excessive deposition of iron in tissues, especially in the liver and pancreas), since iron deposits compromise the production of insulin by the islets of Langerhans in the pancreas. Severe spastic disorders of the muscles in the limbs, head, and face suggests the presence of Wilson's disease, especially if there is a family history, since the copper deposits characteristic of that disorder are toxic to the liver and to structures in the base of the brain. A history of chronic lung infections or of progressive obstructive lung disease may be present in patients with cystic fibrosis or a deficiency of alpha$_1$-antitrypsin.

Congenital disorders producing cirrhosis

Confirmation of cirrhosis

A diagnosis of cirrhosis is confirmed by blood tests that show an elevated concentration of hepatocytic enzymes, reduced levels of coagulation proteins, elevated levels of bilirubin, and, most importantly, reduced amounts of serum albumin (a major protein of human blood plasma) and increases in serum globulin (a specific group of proteins found in blood plasma and including immunoglobulins). Although other tests may also be abnormal in patients with acute liver disease, serum albumin levels are usually not reduced in the acute stage of the disease because that protein is rather long-lived, up to one month, and levels do not decrease until the liver disease becomes chronic. Elevated levels of serum iron or copper support a diagnosis of hemochromatosis or Wilson's disease, respectively, while a positive test for serum antibodies to cellular mitochondria is associated almost solely with primary biliary cirrhosis. The presence of HBV surface antigen or of delta agent suggests posthepatic cirrhosis. A percutaneous needle biopsy of the liver is the most valuable diagnostic test, since this procedure makes available an actual specimen of liver tissue for microscopic examination. Treatment of cirrhosis of the liver never results in a completely normal organ, since the process of scarring and nodular regeneration is permanent. The process itself, however, can be prevented or its progress halted by managing the precipitating factors of the disease.

Complications of advanced liver disease. *Hepatic encephalopathy.* Hepatic encephalopathy refers to the changes in the brain that occur in patients with advanced acute or chronic liver disease. If liver cells are damaged, certain substances that are normally cleansed from the blood by the healthy liver are not removed. In the case of cirrhosis, blood from the portal system is not exposed to functioning hepatocytes because it is transported through blood vessels in the liver that do not run through regenerating nodules of hepatocytes, owing to the atypical growth inherent in the cirrhotic process. These products of cell metabolism are primarily nitrogenous substances derived from protein, especially ammonia, or possibly certain straight-chain fatty acids. They pass to the brain where they damage functioning nervous tissue or subvert the actions of neurotransmitters, chemical messengers that carry impulses from one brain cell to another. In acute diseases, the brain exposed to those agents becomes swollen to the point where normal breathing may cease. Chronic exposure can lead to destruction of nerve cells with replacement by scar tissue (gliosis). A patient with chronic hepatic encephalopathy may develop progressive loss of memory, disorientation, untidiness, and muscular tremors, leading to a form of chronic dementia. The ingestion of protein invariably aggravates these symptoms.

Gliosis

The treatment of hepatic encephalopathy involves, first, the removal of all drugs that require detoxification in the liver and, second, the reduction of the intake of protein. Ammonia is a potentially harmful by-product of digestion, and its concentration in the blood can be lowered either through the reduction of intestinal bacteria by the administration of enteric antibiotics, which reduce the production of ammonia in the colon, or by the administration of lactulose, a nonabsorbable carbohydrate whose by-products make the contents of the colon more acidic, creating an environment that reduces the diffusion of ammonia from the intestinal lumen to the portal blood vessels.

Portal hypertension. Portal hypertension, the increased pressure in the portal vein and its tributaries that is the result of impediments to venous flow into the liver, is brought about by the scarring characteristic of the cirrhotic process. The increased pressure causes feeders of the portal vein to distend markedly, producing varices, or

dilations of the veins. When varices are located in superficial tissues, they may rupture and bleed profusely. Two such locations are the lower esophagus and the perianal region. Esophageal varices are likely to bleed most heavily, and, because of the reduced blood flow in the liver that results and the large amount of protein contained in the blood that is shed into the intestines, profuse bleeding from esophageal varices is frequently associated with the onset of hepatic encephalopathy or coma. Because of their location at the lower end of the esophagus or the upper portion of the stomach, bleeding from varices is often difficult to control. It may stop spontaneously, but it is likely to recur. Considerable success in stemming such hemorrhage and preventing its recurrence has been achieved by the injection of sclerosing (hardening) agents into varices during endoscopic visualization. If variceal bleeding persists and if the patient can withstand a long and complex operative procedure, surgical formation of a shunt, or artificial passageway, from the portal vein or one of its feeders to a systemic abdominal vein, such as the vena cava or the left renal vein, may be done.

Esophageal varices

Ascites. The accumulation of fluid in the abdominal cavity, or ascites, is related to portal hypertension, significant reduction in serum albumin, and renal retention of sodium. When albumin levels in blood are lower than normal, there is a marked reduction in the force that holds plasma water within the blood vessels and normally resists the effects of the intravascular pressure. The resulting increase in intravascular pressure, coupled with the increased internal pressure caused by the portal venous obstruction in the liver, leads to massive losses of plasma water into the abdominal cavity. The associated reduction of blood flow to the kidneys causes increased elaboration of the hormone aldosterone, which, in turn, causes the retention of sodium and water and a reduction in urinary output. In addition, because the movement of intestinal lymph into the liver is blocked by the cirrhotic process in the liver, the backflow of this fluid into the abdominal cavity is greatly increased. The volume of abdominal ascites in adults with cirrhosis may reach levels as great as 10 to 12 litres (10.6 to 12.7 quarts). Ascitic fluid may accumulate in the scrotum and in the chest cavity, where its presence, combined with the upward pressure on the diaphragm from the abdominal fluid, may severely affect breathing. Appetite also is often reduced by the abdominal distention.

Treatment of cirrhotic ascites

The treatment of cirrhotic ascites begins with the removal of enough fluid directly from the abdomen by needle puncture to ease discomfort and breathing. Patients are placed on diets low in salt (sodium chloride), and they are given diuretic drugs to increase the output of water by the kidneys. If these measures do not control massive ascites, ascites can be drained internally into the general venous blood system by running a plastic tube from the abdominal cavity, under the skin of the chest, into the right internal jugular vein of the neck (peritoneovenous shunt of LeVeen).

Hepatorenal syndrome. A progressive reduction in kidney function that often occurs in persons with advanced acute or chronic liver disease, hepatorenal syndrome probably results from an inadequate perfusion of blood through the cortical (outer) portions of the kidneys, where most removal of waste products occurs. In some instances, hepatorenal syndrome is precipitated by marked reductions in blood volume that result from a low concentration of water in the blood. Hemorrhages also can reduce kidney function by leading to damage of renal tubules. Finally, with advanced hepatocytic dysfunction, a spasm of blood vessels in the renal cortex can occur, often with good blood flow to the rest of the kidney. This spasm results in progressive failure in kidney function and often leads to death. The kidneys themselves are frequently undamaged structurally. Treatment of patients with volume depletion and tubular damage often may lead to significant improvement in kidney function.

Tumours. Although not uncommon, cancer originating in the liver, usually in hepatocytes and less frequently in cells of bile duct origin, is rare in the West and is almost always associated with active cirrhosis, particularly the form found in patients with chronic hepatitis. The survival rate from liver cancer is small. In certain underdeveloped countries, especially in tribal Africa, the incidence of this malignancy is high and is a major cause of death in the prevalence of chronic viral hepatitis or the chronic presence of viruses in the blood (viremia) caused by hepatitis B. Long exposure to certain environmental poisons, such as vinyl chloride or carbon tetrachloride, has also been shown to lead to hepatic cancer.

Incidence in underdeveloped countries

Cancers arising elsewhere in the body, particularly in abdominal organs, lungs, and lymphoid tissue, commonly lead to metastatic cancer in the liver and are by far the most frequent type of hepatic malignancy. Usually, when such metastases are found, the primary tumour has advanced beyond the stage where it can be removed surgically. Various benign types of tumours and cysts arise from certain components of the liver, such as the hepatocytes (adenomas) or blood vessels (hemangiomas). While the cause of these lesions is not always clear, hepatic adenomas are associated with the prolonged use of female sex hormones (estrogens). Symptoms of benign tumours depend mainly on their size and their position in relation to the surface of the liver. If they enlarge significantly, patients may note pain or sensations of heaviness in the upper abdomen. When benign tumours are located close to the surface of the liver, they may rupture through the capsule and bleed freely into the abdominal cavity. Surgery is then required.

Benign cysts (tissue swellings filled with fluid) in the liver may occur as congenital defects or as the result of infections from infestation of the dog tapeworm (*Echinococcus granulosus*). Abscesses on the liver result from the spread of infection from the biliary tract or from other parts of the body, especially the appendix and the pelvic organs. Specific liver abscesses also result from infections with the intestinal parasite *Entamoeba histolytica.* Abscesses usually respond well to treatment with specific antibiotics, although surgical drainage is required in some cases.

BILIARY TRACT

Gallstones. Cholelithiasis, or the formation of gallstones in the gallbladder, is the most common disease of the biliary tract. Gallstones are of three types: stones containing primarily calcium bilirubinate (pigment stones); stones containing 25 percent or more of cholesterol; and stones composed of variable mixtures of both bilirubin and cholesterol (mixed gallstones). Purely pigment stones are more common in certain parts of Asia than in the West, and they are prone to occur in persons who suffer from forms of anemia caused by the rapid destruction of red blood cells (hemolysis). Hemolytic disease results from the hereditary or acquired acquisition of abnormal forms of hemoglobin or from abnormalities of the red blood cell membrane in disorders such as sickle-cell anemia, thalassemia, or acquired hemolytic anemias. Increased destruction of red blood cells leads to abnormally large amounts of bilirubin, the hemoglobin derivative, in the liver and the consequent secretion into the biliary tract of increased amounts of the water-soluble conjugate, bilirubin diglucuronide, a pigment that is normally secreted in the urine. In the biliary tract, particularly in the gallbladder, some of this bilirubin diglucuronide is broken down by bacterial or mucosal enzymes into water-insoluble bilirubin, which then tends to form stones. There are two types of pigment stones, black and brown. Black stones tend to form mainly in the gallbladder and occur in sterile bile, while brown stones may occur in any part of the biliary tract in patients with chronic biliary infections and varying degrees of stasis. The reasons for the increased incidence of pigment stones among persons with cirrhosis of the liver and the aged are not clear, although increased red blood cell destruction may play a part. The occurrence of pigment stones is slightly more common in women.

Types of gallstones

Cholesterol and mixed cholesterol-bilirubinate stones occur when the proportion of cholesterol in bile exceeds the capacity of bile acids and lecithin to contain the total amount of cholesterol in micellar colloidal solution. When this critical micellar concentration is surpassed and the

solution is saturated, crystalline particles of cholesterol are formed. The resulting gallstones contain large amounts of crystalline cholesterol and smaller quantities of calcium bilirubinate. Pure cholesterol gallstones are rare.

Risk factors

Cholesterol gallstones occur about twice as frequently in women as they do in men, and at younger ages. Those at increased risk of cholesterol gallstones include persons who are obese, on diets high in caloric content or in cholesterol, diabetic, or taking female sex hormones. Each of these factors favours increased concentrations of cholesterol in bile. In addition, some persons are unable, for genetic reasons, to convert sufficient amounts of cholesterol to bile acids, thus favouring the increased formation of stones. Some illnesses reduce the capacity of the lower small intestine to reabsorb bile acids, leading to deficits of bile acids that cannot be overcome by hepatic synthesis alone. During pregnancy, the ratio of chenodeoxycholic acid to cholic acid in hepatic bile is reduced, thus making bile more prone to produce stones (lithogenic). Decreased flow of bile in the gallbladder, a condition that occurs late in pregnancy, in persons on diets low in fat, and among certain diabetics, also appears to favour the formation of cholesterol stones. Occasionally, some persons produce lithogenic bile, which results from reduced concentrations of phospholipids.

Symptoms are likely to be absent in about half of all patients who have gallstones. When they do appear, symptoms are caused by transient or prolonged obstruction of a portion of the biliary tract, most commonly the cystic duct at the point where it emerges from the gallbladder. This obstruction leads to painful contraction of the gallbladder, swelling of its wall, and acute inflammation (cholecystitis).

Cholecystitis

During an attack of cholecystitis, patients are often found to have fever, sharp pain in the abdomen (which also may be felt in the right shoulder region), tenderness over the region of the gallbladder, and elevations of the white blood cell count. If the obstruction of the neck of the gallbladder is prolonged, bacterial infections may appear, leading to formation of an abscess. Patients with bacterial infections in the gallbladder or bile ducts commonly have severe rigours, or shaking chills, with high, spiking fevers. Jaundice does not occur with gallstone complications unless the stones become impacted and obstruct the common bile duct, thus slowing or interrupting the free passage of bile from the liver to the intestine. This jaundice is associated with a marked lightening of stool colour, caused by the absence of bile pigments in the intestine, and a change in the colour of urine to a dark amber, caused by large quantities of conjugated bilirubin.

Gallstone disease is easy to diagnose since calculi in the gallbladder can be easily detected by ultrasonography. Enlargement of the gallbladder and bile ducts (resulting from obstruction) also can be detected by this method.

As many as one-half of all persons with gallstones never have serious symptoms or complications. Thus, if gallstones are discovered on routine examination or during abdominal surgery for other reasons, and if the patient has no history of gallstone symptoms, nothing probably needs to be done. The situation is different, however, in persons who are clearly symptomatic or who are suffering acute complications, such as cholecystitis or abscesses. The traditional treatment in these cases is surgical removal of the diseased gallbladder and exploration of the bile ducts by X rays at the time of surgery for stones. The risks of this surgery are extremely small, although they do increase considerably in persons with acute complications and in older persons. Once the gallbladder and ductal stones are removed, there is little likelihood that cholesterol or black pigment stones will recur, although brown pigment stones may occasionally recur in the bile ducts after cholecystectomy.

Many cholesterol gallstones can be dissolved without surgery as long as the gallbladder has retained its ability to concentrate bile and the cystic duct is unobstructed. This is accomplished by regular oral administration of the bile acids chenodeoxycholic acid or ursodeoxycholic acid. The ingestion of these acids increases the amount of bile acids in hepatic bile and increases the ratio of bile acids to cholesterol, thus changing the bile from lithogenic to nonlithogenic. This medication must be continued for more than one year for the cholesterol gallstones to be completely dissolved and then continued permanently at reduced doses to prevent the reappearance of stones. Only a small percentage of patients are willing to undergo this permanent treatment, and the use of bile acids is confined either to those who strongly oppose surgery or those for whom surgery imposes great risk. Pigment stones do not respond to bile acid therapy.

Cancer

Other biliary tract disorders. Cancer of the biliary tract is rare but may occur in almost any area, including the gallbladder, the hepatic ducts, the common bile duct, or the ampulla of Vater. About 90 percent of persons with primary cancer of the gallbladder also have gallstones. The risk of cancer in persons with gallstones, however, is very low (about 1 percent or less). In cancer of the bile duct, congenital cysts and parasitic infections, such as liver flukes, seem to lead to increased risks. Persons with extensive chronic ulcerative colitis also show a greater than normal incidence of bile duct carcinoma. Obstructive jaundice is usually the first sign of biliary tract cancer. Surgery is the only treatment, and the chances of cure are very small. Because most biliary duct cancers grow very slowly, physicians often try to relieve the obstructive jaundice by passing tubular stents (supporting devices) through the obstruction, using endoscopic or radiologic techniques.

Postcholecystectomy syndrome comprises painful attacks, often resembling preoperative symptoms, that occasionally occur following the surgical removal of gallstones and the gallbladder. These attacks may be related to intermittent muscular spasms of the sphincter of Oddi or of the bile ducts. Drugs are used to help prevent or reduce these spasms.

Three types of jaundice

Jaundice. Jaundice, or yellowing of the skin, scleras, and mucous membranes, occurs whenever the level of bilirubin in the blood is significantly above normal. This condition is evident in three different types of disorders, more than one of which may be present simultaneously in a single person. The first type, unconjugated, or hemolytic, jaundice, appears when the amount of bilirubin produced from hemoglobin by the destruction of red blood cells or muscle tissue (myoglobin) overwhelms the normal capacity of the liver to transport it or when the ability of the liver to conjugate normal amounts of bilirubin into bilirubin diglucuronide is significantly reduced by inadequate intracellular transport or enzyme systems. The second type, hepatocellular jaundice, arises when liver cells are damaged so severely that their ability to transport bilirubin diglucuronide into the biliary system is reduced, allowing some of this yellow pigment to regurgitate into the bloodstream. The third type, cholestatic, or obstructive jaundice, occurs when essentially normal liver cells are unable to transport bilirubin either through the hepatocytic-bile capillary membrane, because of damage in that area, or through the biliary tract, because of anatomical obstructions (atresias, gallstones, cancer).

Unconjugated jaundice. Unconjugated jaundice, or hemolytic jaundice, is characterized by the absence of bile pigments in the urine and by normal stool colour. The colour of the urine is normal because the bilirubin in the blood is unconjugated to glucouronic acid and therefore bound to blood albumin and insoluble in water. Thus the bilirubin is not filtered by the kidney. The colour of stools remains normal because much of the bilirubin in the blood is filtered normally by the liver and enters the intestine promptly by way of the biliary system. Hemolytic diseases in newborn infants may lead to serious brain damage (kernicterus) if the unconjugated bilirubin crosses into the brain stem and destroys vital nuclei. The exposure to blue light of infants at risk for kernicterus converts the bilirubin to harmless and colourless degradation products. Unconjugated hyperbilirubinemia also occurs in many newborn infants, especially if they are premature, when the bilirubin transport enzyme systems are not fully developed (physiologic jaundice of the newborn). This disorder is self-limited, may require occasional exposures to blue light, and usually disappears within the first two weeks of extrauterine life. In Gilbert's disease, there is a fairly common hereditary reduction in one hepatic transport

Effects on newborn infants

protein, ligandin, and one conjugating enzyme, glucuronyl transferase. This results in a harmless lifelong tendency to mild degrees of unconjugated jaundice, especially during periods of fasting or fatigue.

Hepatocellular jaundice. A characteristic in all types of hepatitis and cirrhosis and in congestive liver disease, hepatocellular jaundice is characterized by dark amber urine and normal or slightly paler than normal stools. In hepatocellular jaundice, because much of the bilirubin in the blood already has been conjugated by the endoplasmic reticulum of the hepatocyte, it is water-soluble and can be filtered by the kidney. Stools are usually normal because some bile pigment also manages to be excreted into the biliary tract and intestine.

Cholestatic jaundice. Cholestatic jaundice is also distinguished by amber-coloured urine, but the colour of the stools is likely to be very pale (clay-coloured) owing to the failure of bile pigments to pass into the intestine. Itching of the skin is commonly associated with this condition. Cholestasis occurs in many types of hepatitis, especially those caused by certain drugs, and in diseases that primarily damage small bile passages in the liver (intrahepatic cholestasis). Cholestatic jaundice also occurs in patients with obstructive disorders of the biliary tract outside of the liver (extrahepatic cholestasis). It is often impossible to determine the level of obstruction by means of examination alone, and more sophisticated testing is required to locate the site of damage.

PANCREAS

Pancreatitis. Inflammation of the pancreas, or pancreatitis, is probably the most common disease of this organ. The disorder may be confined to either singular or repeated acute episodes, or it may become a chronic disease. There are many factors associated with the onset of pancreatitis, including direct injury, certain drugs, viral infections, heredity, hyperlipidemia (increased levels of blood fats), and congenital derangements of the ductal system. In Western society most cases are related either to alcoholism or to gallstones, especially when stones pass spontaneously into the ampulla of Vater. Although the immediate cause of acute pancreatitis is not always clear, it seems to involve one or more of the following factors: heavy stimulation of pancreatic acini; increased pressure within the duct because of partial obstruction (gallstones) or edema (alcohol); and damage to the fine ductal network in the gland, which allows the escape of activated, potent, and destructive digestive enzymes into the substance of the pancreas itself and into surrounding tissues. Overstimulation of mechanisms of secretory enzyme production in the acinar cell may also lead to the energizing of intracellular (lysosomal) enzyme systems, resulting in the conversion of proenzymes to active forms that begin to digest cellular organelles. The gland thus begins to self-destruct. Similar damage may appear in other body organs, such as the lungs, kidneys, and blood vessels, which receive these activated enzymes by way of the bloodstream. It is not clear how the proenzyme trypsinogen is converted to trypsin in the damaged acinar cell, but it is known that the activation of the other proenzymes proceeds from this conversion. The extent of acinar destruction appears to depend on the strength of the causative factors.

Factors in the onset of pancreatitis

Localized, severe abdominal and midback pain resulting from enzyme leakage, tissue damage, and nerve irritation is the most common symptom of acute pancreatitis. In severe cases, respiratory failure, shock, and even death may occur. The severity of the symptoms generally depends on the extent of the damage to the pancreas; the mortality rate approaches 50 percent in severe (hemorrhagic) pancreatitis but is less than 5 percent in milder forms. The diagnosis is confirmed by the detection of elevated levels of pancreatic enzymes (amylase and lipase) in the blood and, if islet cell function is disturbed by the inflammatory process, elevated blood glucose values. Ultrasonographic or computed tomographic scans of the upper abdomen usually reveal an enlarged and swollen pancreas. Sustained pain, often with fever, suggests the presence of a pseudocyst or abscess caused by localized areas of destruction and infections in the pancreas.

Mortality rate

Acute pancreatitis is treated primarily by supportive therapy, with replacement of fluid and salt and control of pain. In severe cases, washing necrotic material and active enzymes from the abdominal cavity during surgery may be beneficial. Following recovery from an acute attack, the prevention of further attacks should be the primary goal. Thus, the removal of gallstones, cessation of alcohol ingestion, lowering of blood fats through diet, and discontinuation of toxic drugs (glucocorticoids and thiazide diuretics, for example) can be helpful measures. In instances where repeated attacks of acute pancreatitis have resulted in strictures (scars) of the main pancreatic duct, surgical repair may prevent further attacks.

Chronic pancreatitis. Chronic pancreatitis rarely follows repeated acute attacks. It seems instead to be a separate disorder that results in mucus plugs and precipitation of calcium salts in the smaller pancreatic ducts. The progressive loss of acinar and islet cell function follows, presumably as a consequence of continuous inflammation resulting from the ductal blockage. Progressive calcification, which at times results in the formation of large stones in the major pancreatic ducts, has been attributed to diminished production of an acinar protein that normally holds calcium in solution. Alcoholism and certain hereditary factors account for almost all of the cases of chronic pancreatitis seen in Western countries. Chronic protein malnutrition is an important element in underdeveloped countries. Recurrent abdominal pain, diabetes, and intestinal malabsorption of dietary nutrients are the main symptoms of chronic pancreatitis. Weight loss and deficiencies of fat-soluble vitamins (A, D, E, and K) are common. Treatment includes abstinence from alcohol, management of diabetes with insulin, and ingestion of pancreatic enzyme supplements to control dietary malabsorption.

Symptoms

Cystic fibrosis of the pancreas. Cystic fibrosis is inherited, but it is not expressed unless both members of a pair of hemologous, or corresponding, chromosomes carry the trait. The major functional abnormality in persons with the disease appears to be the elaboration by mucous glands throughout the body of secretions containing greater than normal concentrations of protein and calcium. This imbalance leads to increased viscosity of the secretions and precipitation of mucus and organic constituents in gland ducts. The resulting plugging process in the pancreas almost invariably causes destruction and scarring of the acinar tissue, usually without damaging the islets of Langerhans. A similar process in the hepatic biliary system produces foci of fibrosis and bile duct proliferation, a singular form of cirrhosis. In cystic fibrosis, the resulting pancreatic insufficiency usually can be treated by the oral replacement of pancreatic enzymes.

Cancer. Carcinoma of the pancreas arises primarily from the ductal system. The incidence of carcinoma of the pancreas has increased slightly (somewhat more in men than in women) and now exceeds cancer of the stomach. No certain risk factors have been identified, although suggestions have been made that pancreatic cancer occurs at increased rates among diabetics and persons with chronic pancreatitis. Upper abdominal pain, often radiating to the back, and weight loss are the most common symptoms of pancreatic cancer. Obstructive jaundice is a frequent symptom when the head of the pancreas is involved. The diagnosis is readily made in most cases by computed tomographic examination, at times supplemented by biopsy. There is no effective treatment, and more than 90 percent of patients die within the first year after diagnosis. If the tumour mass is localized and has not invaded blood vessels and nerves surrounding the pancreas, it occasionally can be removed surgically. Jaundice and intestinal obstruction can be relieved temporarily by surgical bypass procedures. Radiation and chemotherapy have shown some promise as therapeutic agents if they are started promptly in the course of the disease and continued for long periods.

(H.J.Dw.)

BIBLIOGRAPHY

General features of digestion and absorption: WILLIAM T. KEETON, JAMES L. GOULD, and CAROL GRANT GOULD, *Biological Science*, 4th ed. (1986), is a comprehensive study that includes an examination of digestion. CHARLES J. FLICKINGER *et al.*,

Medical Cell Biology (1979); and WALTER HOPPE *et al.* (eds.), *Biophysics* (1983; originally published in German, 1977), examine the structure of cells and cell membranes, the molecular mechanics of peptides, and the function of enzymes. Specialized studies include H.J. VONK and J.H.R. WESTERN, *Comparative Biochemistry and Physiology of Enzymatic Digestion* (1984); THOMAS H. WILSON, *Intestinal Absorption* (1962); D.H. SMYTH (ed.), *Intestinal Absorption*, 2 vol. (1974); B.P. BABKIN, *Secretory Mechanisms of the Digestive Glands*, 2nd rev. ed. (1950); A.S.V. BURGEN and N.G. EMMELIN, *Physiology of the Salivary Glands* (1961); ARNOLD V. WOLF, *Thirst* (1958); GEOFFREY H. BOURNE and GEORGE W. KIDDER (eds.), *Biochemistry and Physiology of Nutrition*, 2 vol. (1953); E.F. ANNISON and D. LEWIS, *Metabolism in the Rumen* (1959); and J.B. JENNINGS, *Feeding, Digestion, and Assimilation in Animals*, 2nd ed. (1972).

Anatomy and physiology of the digestive system: Chapters on digestion can be found in such comprehensive texts as HENRY GRAY, *Anatomy of the Human Body*, 30th American ed., edited by CARMINE D. CLEMENTE (1985); B.I. BALINSKY, *An Introduction to Embryology*, 5th ed. (1981); and LESLIE BRAINERD AREY, *Developmental Anatomy*, 7th rev. ed. (1974). FRANK H. NETTER, *The Digestive System: A Compilation of Paintings on the Normal and Pathologic Anatomy*, 3 vol. (1957–62), is a part of the Ciba Collection of Medical Illustration series. Specialized works include LEONARD R. JOHNSON *et al.* (eds.), *Physiology of the Gastrointestinal Tract*, 2nd ed. (1981); HORACE W. DAVENPORT, *Physiology of the Digestive Tract*, 5th ed. (1982); CHARLES F. CODE (ed.), *Alimentary Canal*, 5 vol. (1967–68); JOHN MORTON, *Guts: The Form and Function of the Digestive System*, 2nd ed. (1979); R.J. LAST, *Anatomy, Regional and Applied*, 7th ed. (1984); ALFRED SHERWOOD ROMER and THOMAS S. PARSONS, *The Vertebrate Body*, 6th ed. (1986); C. LADD PROSSER (ed.), *Comparative Animal Physiology*, 3rd ed. (1973); J.A. COLIN NICOL, *The Biology of Marine Animals*, 2nd ed. (1967); and ROBERT D. BARNES, *Invertebrate Zoology*, 4th ed. (1980).

Disorders and diseases of the digestive system: Gastrointestinal diseases are treated in such works as CHARLES H. BEST, *Best and Taylor's Physiological Basis of Medical Practice*, 11th ed., edited by JOHN B. WEST (1985); and E.J. HOLBOROW and W.G. REEVES (eds.), *Immunology in Medicine: A Comprehensive Guide to Clinical Immunology*, 2nd ed. (1983). Specialized studies include HARVEY J. DWORKEN, *Gastroenterology: Pathophysiology and Clinical Applications* (1982); MARVIN H. SLEISENGER and JOHN S. FORDTRAN (eds.), *Gastrointestinal Disease: Pathophysiology, Diagnosis, Management*, 3rd ed. (1983); DAVID J.C. SHEARMAN and NIALL D.C. FINLAYSON, *Diseases of the Gastrointestinal Tract and Liver* (1982); H.L. DUTHIE (ed.), *Gastrointestinal Motility in Health and Disease* (1978); F. AVERY JONES, J.W.P. GUMMER, and J.E. LENNARD-JONES, *Clinical Gastroenterology*, 2nd ed. (1968); MOSES PAULSON (ed.), *Gastroenterologic Medicine* (1969); HENRY L. BOCKUS, *Bockus Gastroenterology*, 4th ed., edited by J. EDWARD BERK *et al.*, 7 vol. (1985); and BRIAN M. BARKER and DAVID A. BENDER (eds.), *Vitamins in Medicine*, 4th ed., 2 vol. (1980–82). For current research in the field, see *Gastroenterology Annual*.

(W.T.Ke./N.C.H./W.S./H.J.Dw.)

Dinosaurs

Dinosaur is the common name given to a group of reptiles, often very large, that first appeared at least 228 million years ago, during the Late Triassic Period, and thrived worldwide for more than 150 million years. Most died out by the end of the Cretaceous Period, but many lines of evidence now show that one lineage evolved into modern-day birds.

The name dinosaur comes from the Greek words *deinos* ("terrible" or "fearfully great") and *sauros* ("reptile" or "lizard"). The English anatomist Richard Owen proposed the formal term "Dinosauria" in 1842 to include three giant extinct animals (*Megalosaurus, Iguanodon,* and *Hylaeosaurus*) represented by large fossilized bones that had been unearthed at several locations in southern England during the early part of the 19th century. Owen recognized that these reptiles were far different from other known reptiles of the present and the past for three reasons: they were large yet obviously terrestrial, unlike the aquatic ichthyosaurs and plesiosaurs that were already known; they had five vertebrae in their hips, whereas most known reptiles have only two; and, rather than holding their limbs sprawled out to the side in the manner of lizards, dinosaurs held their limbs under the body in columnar fashion, like elephants and other large mammals.

Originally applied to just a handful of incomplete specimens, the category Dinosauria now encompasses more than 800 generic names and at least 1,000 species, with new names being added to the roster every year as the result of scientific explorations around the world. Not all of these names are valid taxa, however. A great many of them have been based on fragmentary or incomplete material that may actually have come from two or more different dinosaurs. In addition, bones have sometimes been misidentified as dinosaurian when they are not from dinosaurs at all. Nevertheless, dinosaurs are well documented by abundant fossil remains recovered from every continent on Earth, and the number of known dinosaurian taxa is estimated to be 10–25 percent of actual past diversity.

The extensive fossil record of genera and species is testimony that dinosaurs were diverse animals with widely varying lifestyles and adaptations. Their remains are found in sedimentary rock layers (strata) dating to the Late Triassic Period (230 million to 208 million years ago). The abundance of their fossilized bones is substantive proof that dinosaurs were the dominant form of terrestrial animal life during the Mesozoic Era (245 million to 66.4 million years ago). It is likely that the known remains represent a very small fraction (probably less than 0.0001 percent) of all the individual dinosaurs that once lived.

For coverage of related topics in the *Macropædia* and *Micropædia,* see the *Propædia,* section 313, and the *Index.*

This article is divided into the following sections:

The search for dinosaurs 315
  The first finds 315
  Reconstruction and classification 316
  American hunting expeditions 316
  Dinosaur ancestors 317
  Modern studies 317
  Extinction 317
    Faunal changes
    The K–T boundary event
    The asteroid theory
    Dinosaur descendants
Natural history 319
  Habitats 319
  Food and feeding 319
    The plant eaters
    The flesh eaters
  Herding behaviour 319
  Growth and life span 320
  Reproduction 320
  Body temperature 321
    Ectothermy and endothermy
    Clues to dinosaurian metabolism
Classification 322
  Saurischia 322
    Sauropodomorpha
      Prosauropoda
      Sauropoda
    Theropoda
      Ceratosauria
      Tetanurae
  Ornithischia 326
    Cerapoda
      Ornithopoda
      Pachycephalosauria
      Ceratopsia
    Thyreophora
      Stegosauria
      Ankylosauria
Bibliography 330

The search for dinosaurs

THE FIRST FINDS

Before Richard Owen introduced the term "Dinosauria" in 1842, there was no concept of anything even like a dinosaur. Large fossilized bones quite probably had been observed long before that time, but there is little record—and no existing specimens—of such findings much before 1818. In any case, people could not have been expected to understand what dinosaurs were even if they found their remains. For example, some classical scholars now conclude that the Greco-Roman legends of griffins from the 7th century BC were inspired by discoveries of protoceratopsian dinosaurs in the Altai region of Mongolia. In 1676 Robert Plot of the University of Oxford included, in a work of natural history, a drawing of what was apparently the knee-end of the thigh bone of a dinosaur, which he thought might have come from an elephant taken to Britain in Roman times. Fossil bones of what were undoubtedly dinosaurs were discovered in New Jersey in the late 1700s and were probably discussed at the meetings of the American Philosophical Society in Philadelphia. Soon thereafter, Lewis and Clark's expedition encountered dinosaur fossils in the western United States.

Initial explanations

The earliest verifiable published record of dinosaur remains that still exists is a note in the 1820 *American Journal of Science and Arts* by Nathan Smith. The bones described had been found in 1818 by Solomon Ellsworth, Jr., while he was digging a well at his homestead in Windsor, Connecticut. At the time, the bones were thought to be human, but much later they were identified as *Anchisaurus.* Even earlier (1800), large birdlike footprints had been noticed on sandstone slabs in Massachusetts. Pliny Moody, who discovered these tracks, attributed them to "Noah's raven," and Edward Hitchcock of Amherst College, who began collecting them in 1835, considered them to be those of some giant extinct bird. The tracks are now recognized as having been made by several different kinds of dinosaurs, and such tracks are still commonplace in the Connecticut River Valley today.

Better known are the finds in southern England during the early 1820s by William Buckland (a clergyman) and Gideon Mantell (a physician), who described *Megalosaurus* and *Iguanodon,* respectively. In 1824 Buckland

published a description of *Megalosaurus,* fossils of which consisted mainly of a lower jawbone with a few teeth. The following year Mantell published his "Notice on the *Iguanodon,* a Newly Discovered Fossil Reptile, from the Sandstone of Tilgate Forest, in Sussex," on the basis of several teeth and some leg bones. Both men collected fossils as an avocation and are credited with the earliest published announcements in England of what later would be recognized as dinosaurs. In both cases their finds were too fragmentary to permit a clear image of either animal. In 1834 a partial skeleton was found near Brighton that corresponded with Mantell's fragments from Tilgate Forest. It became known as the Maidstone *Iguanodon,* after the village where it was discovered. The Maidstone skeleton provided the first glimpse of what these creatures might have looked like.

Two years before the Maidstone *Iguanodon* came to light, a different kind of skeleton was found in the Weald of southern England. It was described and named *Hylaeosaurus* by Mantell in 1832 and later proved to be one of the armoured dinosaurs. Soon others began finding fossil bones in Europe. Owen identified two additional dinosaurs, albeit from fragmentary evidence: *Cladeiodon,* which was based on a single large tooth, and *Cetiosaurus,* which he named from an incomplete skeleton composed of very large bones. Having carefully studied most of these fossil specimens, Owen recognized that all of these bones represented a group of large reptiles that were unlike any living varieties. In a report to the British Association for the Advancement of Science in 1841, he described these animals, and the word "Dinosauria" was first published in the association's proceedings in 1842.

RECONSTRUCTION AND CLASSIFICATION

During the decades that followed Owen's announcement, many other kinds of dinosaurs were discovered and named in England and Europe: *Massospondylus* in 1854, *Scelidosaurus* in 1859, *Bothriospondylus* in 1875, and *Omosaurus* in 1877. Popular fascination with the giant reptiles grew, reaching a peak in the 1850s with the first attempts to reconstruct the three animals on which Owen based Dinosauria—*Iguanodon, Megalosaurus,* and *Hylaeosaurus*—for the first world exposition, the Great Exhibition of 1851 in London's Crystal Palace. A sculptor under Owen's direction (Waterhouse Hawkins) created life-size models of these two genera, and in 1854 they were displayed together with models of other extinct and living reptiles, such as plesiosaurs, ichthyosaurs, and crocodiles.

By the 1850s it had become evident that the reptile fauna of the Mesozoic Era was far more diverse and complex than it is today. The first important attempt to establish an informative classification of the dinosaurs was made by the English biologist T.H. Huxley as early as 1868. Because he observed that these animals had legs similar to birds as well as other birdlike features, he established a new order called Ornithoscelida. He divided the order into two suborders. Dinosauria was the first and included the iguanodonts, the large carnivores (or megalosaurids), and the armoured forms (including *Scelidosaurus*). Compsognatha was the second order, named for the very small birdlike carnivore *Compsognathus.*

Two types of dinosaurs

Huxley's classification was replaced by a radically new scheme proposed in 1887 by his fellow Englishman H.G. Seeley, who noticed that all dinosaurs possessed one of two distinctive pelvic designs, one like that of birds and the other like that of reptiles. Accordingly, he divided the dinosaurs into the orders Ornithischia (having a birdlike pelvis) and Saurischia (having a reptilian pelvis). Ornithischia included four suborders: Ornithopoda (*Iguanodon* and similar herbivores), Stegosauria (plated forms), Ankylosauria (*Hylaeosaurus* and other armoured forms), and Ceratopsia (horned dinosaurs, just then being discovered in North America). Seeley's second order, the Saurischia, included all the carnivorous dinosaurs, such as *Megalosaurus* and *Compsognathus,* as well as the giant herbivorous sauropods, including *Cetiosaurus* and several immense "brontosaur" types that were turning up in North America. In erecting Saurischia and Ornithischia, Seeley cast doubt on the idea that Dinosauria was a natural grouping of these animals. This uncertainty persisted for a century thereafter, but it is now understood that the two groups share unique features that indeed make the Dinosauria a natural group.

By courtesy of the Institut Royal des Sciences Naturelles de Belgique, Brussels

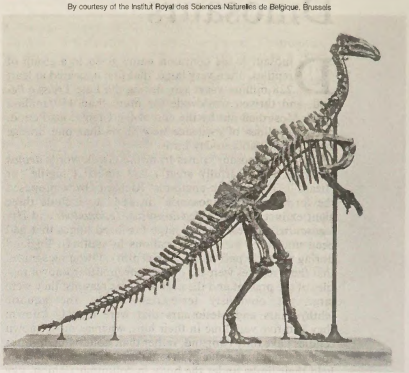


Figure 1: *Iguanodon* skeleton reconstructed with an upright posture by Louis Dollo in the 19th century.

A herd of *Iguanodon*

In 1878 a spectacular discovery was made in the town of Bernissart, Belgium, where several dozen complete articulated skeletons of *Iguanodon* were accidentally uncovered in a coal mine during the course of mining operations. Under the direction of the Royal Institute of Natural Science of Belgium, thousands of bones were retrieved and carefully restored over a period of many years. The first skeleton was placed on exhibit in 1883, and today the public can view an impressive herd of *Iguanodon.* The discovery of these multiple remains gave the first hint that at least some dinosaurs may have traveled in groups and showed clearly that some dinosaurs were bipedal (walking on two legs). The supervisor of this extraordinary project was Louis Dollo, a zoologist who was to spend most of his life studying *Iguanodon,* working out its structure, and speculating on its living habits (Figure 1).

AMERICAN HUNTING EXPEDITIONS

England and Europe produced most of the early discoveries and students of dinosaurs, but North America soon began to contribute a large share of both. One leading student of fossils was Joseph Leidy of the Academy of Natural Sciences in Philadelphia, who named some of the earliest dinosaurs found in America, including *Palaeoscincus, Trachodon, Troodon,* and *Deinodon.* Unfortunately, some names given by Leidy are no longer used, because they were based on such fragmentary and undiagnostic material. Leidy is perhaps best known for his study and description of the first dinosaur skeleton to be recognized in North America, that of a duckbill, or hadrosaur, found at Haddonfield, New Jersey, in 1858, which he named *Hadrosaurus foulkii.* Leidy's inference that this animal was probably amphibious influenced views of dinosaur life for the next century.

Two Americans whose work during the second half of the 19th century had worldwide impact on the science of paleontology in general, and the growing knowledge of dinosaurs in particular, were O.C. Marsh of Yale College and E.D. Cope of Haverford College, the University of Pennsylvania, and the Academy of Natural Sciences in Philadelphia. All previous dinosaur remains had been discovered by accident in well-populated regions with temperate, moist climates, but Cope and Marsh astutely focused their attention on the wide arid expanses of bare exposed rock in western North America. In their intense quest to find and name new di-

nosaurs, these scientific pioneers became fierce and unfriendly rivals.

Marsh's field parties explored widely, exploiting dozens of now famous areas, among them Yale's sites at Morrison and Canon City, Colorado, and, most important, Como Bluff in southeastern Wyoming. The discovery of Como Bluff in 1877 was a momentous event in the history of paleontology that generated a burst of exploration and study as well as widespread public enthusiasm for dinosaurs. Como Bluff brought to light one of the greatest assemblages of dinosaurs, both small and gigantic, ever found. For decades the site went on producing the first known specimens of Late Jurassic Period (163 million to 144 million years ago) dinosaurs such as *Stegosaurus, Camptosaurus, Camarasaurus, Laosaurus, Coelurus,* and others. From the Morrison site came the original specimens of *Allosaurus, Diplodocus, Atlantosaurus,* and *Brontosaurus* (later renamed *Apatosaurus*). Canon City provided bones of a host of dinosaurs, including *Stegosaurus, Brachiosaurus, Allosaurus,* and *Camptosaurus.* Marsh's specimens now form the core of the Mesozoic collections at the National Museum of Natural History of the Smithsonian Institution and the Peabody Museum of Natural History at Yale University.

Cope's dinosaur explorations ranged as far as, or farther than, Marsh's, and his interests encompassed a wider variety of fossils. Owing to a number of circumstances, however, Cope's dinosaur discoveries were fewer and his collections far less complete than those of Marsh. Perhaps his most notable achievement was finding and proposing the names for *Coelophysis* and *Monoclonius.* Cope's dinosaur explorations began in the eastern badlands of Montana, where he discovered *Monoclonius* in the Judith River Formation of the Cretaceous Period (97.5 million to 66.4 million years ago). Accompanying him there was a talented young assistant, Charles H. Sternberg. Later Sternberg and his three sons went on to recover countless dinosaur skeletons from the Oldman and Edmonton formations of the Late Cretaceous along the Red Deer River of Alberta, Canada.

DINOSAUR ANCESTORS

During the early decades of dinosaur discoveries, little thought was given to their evolutionary ancestry. Not only were the few specimens known unlike any living animal, but they were so different from any other reptiles that it was difficult to discern much about their relationships. Early on it was recognized that, as a group, dinosaurs appear to be most closely allied to crocodilians, though T.H. Huxley had proposed in the 1860s that dinosaurs and birds must have had a very close common ancestor in the distant past. Three anatomic features—socketed teeth, a skull with two large holes (diapsid), and another hole in the lower jaw—are present in both crocodiles and dinosaurs. The earliest crocodilians occurred nearly simultaneously with the first known dinosaurs, so neither could have given rise to the other. It was long thought that the most likely ancestry of dinosaurs could be found within a poorly understood group of Triassic reptiles termed thecodontians ("socket-toothed reptiles"). Today it is recognized that "thecodontian" is simply a name for the basal, or most primitive, members of the archosaurs ("ruling reptiles"), a group that is distinguished by the three anatomic features mentioned above and that includes dinosaurs, pterosaurs (flying reptiles), crocodiles, and their extinct relatives. An early candidate for the ancestor of dinosaurs was a small basal archosaur from the Early Triassic Period (245 million to 240 million years ago) of South Africa called *Euparkeria.* New discoveries suggest creatures that are even more dinosaur-like from the Middle Triassic (240 million to 230 million years ago) and from an early portion of the Late Triassic (230 million to 208 million years ago) of South America; these include *Lagerpeton, Lagosuchus, Pseudolagosuchus,* and *Lewisuchus.* Other South American forms such as *Eoraptor* and *Herrerasaurus* are particularly dinosaurian in appearance and are sometimes considered dinosaurs.

An uncertain origin

The earliest appearance of "true dinosaurs" is almost impossible to pinpoint, since it can never be known with certainty whether the very first (or last) specimen of any kind of organism has been found. The succession of deposits containing fossils is discontinuous and contains many gaps; even within these deposits, the fossil record of dinosaurs and other creatures contained within is far from complete. Further complicating matters is that evolution from ancestral to descendant form is usually a stepwise process. Consequently, as more and more gaps are filled between the first dinosaurs and other archosaurs, the number of features distinguishing them becomes smaller and smaller. Currently, paleontologists define dinosaurs as *Triceratops* (representing Ornithischia), birds (the most recent representatives of the Saurischia), and all the descendants of their most recent common ancestor. Compared with most of their contemporaries, dinosaurs had an improved stance and posture with a resulting improved gait and, in several independent lineages, an overall increase in size. They also were more efficient at gathering food and processing it and apparently had higher metabolic rates and cardiovascular nourishment. All these trends, individually or in concert, probably contributed to the collective success of dinosaurs, which resulted in their dominance among the terrestrial animals of the Mesozoic.

MODERN STUDIES

During the first century or more of dinosaur awareness, workers in the field more or less concentrated on the search for new specimens and new types. Their discoveries then required detailed description and analysis, followed by comparisons with other known dinosaurs in order to classify the new finds and develop hypotheses about evolutionary relationships. These pursuits continue, but newer methods of exploration and analysis have been adopted. Emphasis has shifted from purely descriptive procedures to analyses of relationships by using the methods of cladistics, which dispenses with the traditional taxonomic hierarchy in favour of "phylogenetic trees" that are more explicit about evolutionary relationships. Phylogenetic analyses also help us to understand how certain features evolved in groups of dinosaurs and give us insight into their possible functions.

Reconstructing dinosaurs

Functional anatomic studies extensively use analogous traits of present-day animals that, along with both mechanical and theoretical models, make it possible to visualize certain aspects of extinct animals. For example, estimates of normal walking and maximum running speeds can be calculated on the basis of the analysis of trackways, which can then be combined with biomechanical examination of the legs and joints and reconstruction of limb musculature. Similar methods have been applied to jaw mechanisms and tooth wear patterns to obtain a better understanding of feeding habits and capabilities.

The soft parts of dinosaurs are only imperfectly known. Original colours and patterns cannot be known, but skin textures have occasionally been preserved. Most show a knobby or pebbly surface rather than a scaly texture as in most living reptiles. Impressions of internal organs are rarely preserved, but, increasingly, records of filaments and even feathers have been found on some dinosaurs. Gastroliths ("stomach stones") used for processing food in the gizzard have been recovered from a variety of dinosaurs.

EXTINCTION

A misconception commonly portrayed in popular books and media is that all the dinosaurs died out at the same time—and apparently quite suddenly—at the end of the Cretaceous Period 66.4 million years ago. This is not entirely correct, and not only because birds are a living branch of dinosaurian lineage. The best records, which are almost exclusively from North America, show that dinosaurs were already in decline during the latest portion of the Cretaceous. The causes of this decline, as well as the fortunes of other groups at the time, are complex and difficult to attribute to a single source. In order to understand extinction, it is necessary to understand the basic fossil record of dinosaurs.

Faunal changes. During the 180 million years or so of the Mesozoic Era (245 million to 66.4 million years ago) from which dinosaurs are known, there were constant changes in dinosaur communities. Different species

evolved rapidly and were quickly replaced by others throughout the Mesozoic; it is rare that any particular type of dinosaur survived from one geologic formation into the next, but the overall picture is quite clear: throughout Mesozoic time there was a continuous dying out and renewal of dinosaurian life.

Extinction processes

It is important to note that extinction is a normal, universal occurrence. Mass extinctions often come to mind when the term extinction is mentioned, but the normal background extinctions that occur throughout geologic time probably account for most losses of biodiversity. Just as new species constantly split from existing ones, existing species are constantly becoming extinct. The speciation rate of a group must, on balance, exceed the extinction rate in the long run, or that group will become extinct. The history of animal and plant life is replete with successions as early forms are replaced by new and often more advanced forms. In most instances the layered (stratigraphic) nature of the fossil record gives too little information to show whether the old forms were actually displaced by the new successors (from the effects of competition, predation, or other ecological processes) or if the new kinds simply expanded into the declining population's ecological niches.

Because the fossil record is episodic rather than continuous, it is very useful for asking many kinds of questions, but it is not possible to say precisely how long most dinosaur species or genera actually existed. Moreover, because the knowledge of the various dinosaur groups is somewhat incomplete, the duration of any particular dinosaur can be gauged only approximately—usually by stratigraphic boundaries and presumed "first" and "last" occurrences. The latter often coincide with geologic age boundaries; in fact, the absence of particular life-forms has historically defined most geologic boundaries ever since the geologic record was first compiled and analyzed in the late 18th century. The "moments" of apparently high extinction levels among dinosaurs were near the ends of two stages of the Triassic (about 225 million and 208 million years ago), perhaps at the end of the Jurassic (144 million years ago), and of course at the end of the Cretaceous (66.4 million years ago). Undoubtedly, there were lesser extinction peaks at other times in between, but there are poor terrestrial records for most of the world in the Middle Triassic, Middle Jurassic, and mid-Cretaceous.

The K–T boundary event. It was not only the dinosaurs that disappeared 66.4 million years ago at the Cretaceous–Tertiary, or K–T, boundary. Many other organisms became extinct or were greatly reduced in abundance and diversity, and the extinctions were quite different between, and even among, marine and terrestrial organisms. Land plants did not respond in the same way as land animals, and not all marine organisms showed the same patterns of extinction. Some groups died out well before the K–T boundary, including flying reptiles (pterosaurs) and sea reptiles (plesiosaurs, mosasaurs, and ichthyosaurs). Strangely, turtles, crocodilians, lizards, and snakes were either not affected or affected only slightly. Effects on amphibians and mammals were mild. These patterns seem odd, considering how environmentally sensitive and habitat-restricted many of these groups are today.

Whatever factors caused it, there was undeniably a major, worldwide biotic change near the end of the Cretaceous. But the extermination of the dinosaurs is the best-known change by far, and it has been a puzzle to paleontologists, geologists, and biologists for two centuries. Many hypotheses have been offered over the years to explain dinosaur extinction, but only a few have received serious consideration. Proposed causes have included everything from disease to heat waves and resulting sterility, freezing cold spells, the rise of egg-eating mammals, and X rays from a nearby exploding supernova. Since the early 1980s, attention has focused on the so-called asteroid theory put forward by the American geologist Walter Alvarez, his father, physicist Luis Alvarez, and their coworkers. This theory is consistent with the timing and magnitude of some extinctions, especially in the oceans, but it does not fully explain the patterns on land and does not eliminate the possibility that other factors were at work on land as well as in the seas.

One important question is whether the extinctions were simultaneous and instantaneous or whether they were nonsynchronous and spread over a long time. The precision with which geologic time can be measured leaves much to be desired no matter what means are used (radiometric, paleomagnetic, or the more traditional measuring of fossil content of stratigraphic layers). Only rarely does an "instantaneous" event leave a worldwide—or even regional—signature in the geologic record in the way that a volcanic eruption does locally. Attempts to pinpoint the K–T boundary event, even by using the best radiometric dating techniques, result in a margin of error on the order of 50,000 years. Consequently, the actual time involved in this, or any of the preceding or subsequent extinctions, has remained undetermined.

The asteroid theory. The discovery of an abnormally high concentration of the rare metal iridium at, or very close to, the K–T boundary provides what has been recognized as one of those rare instantaneous geologic time markers that seem to be worldwide. This iridium anomaly, or spike, was first found by Walter Alvarez in the Cretaceous–Tertiary stratigraphic sequence at Gubbio, Italy, in the 1970s. The spike has subsequently been detected at hundreds of localities in Denmark and elsewhere, both in rock outcrops on land and in core samples drilled from ocean floors.

Because the levels of iridium are higher in meteorites than on the Earth, the Gubbio anomaly is thought to have an extraterrestrial explanation. If this is true, such extraterrestrial signatures will have a growing influence on the precision with which geologic time boundaries can be specified. The level of iridium in meteorites has been accepted as representing the average level throughout the solar system and, by extension, the universe. Accordingly, the iridium concentration at the K–T boundary is widely attributed to a collision between the Earth and a huge meteor or asteroid. The impact site (called an astrobleme) of such a collision may have been identified in the Chicxulub crater in Mexico's Yucatán Peninsula.

The asteroid theory is widely accepted as the most probable explanation of the K–T iridium anomaly, but it does not appear to account for all the paleontological data. An impact explosion of this kind would have ejected an enormous volume of terrestrial and asteroid material into the atmosphere, producing a cloud of dust and solid particles that would have encircled the Earth and blocked out sunlight for many months, possibly years. The loss of sunlight could have eliminated photosynthesis and resulted in the death of plants and the subsequent extinction of herbivores, their predators, and scavengers.

Non-asteroid theories

The K–T mass extinctions, however, do not seem to be fully explained by this hypothesis. The stratigraphic record is most complete for extinctions of marine life—foraminifera, ammonites, coccolithophores, and the like. These apparently died out suddenly and simultaneously, and their extinction accords best with the asteroid theory. The fossil evidence of land dwellers, however, suggests a gradual rather than a sudden decline in dinosaurian diversity (and possibly abundance). Alterations in terrestrial life seem to be best accounted for by environmental factors, such as the consequences of seafloor spreading and continental drift, resulting in continental fragmentation, climatic deterioration, increased seasonality, and perhaps changes in the distributions and compositions of terrestrial communities. But one phenomenon does not preclude another. It is entirely possible that a culmination of ordinary biological changes and some catastrophic events, including increased volcanic activity, took place around the end of the Cretaceous.

Dinosaur descendants. Contrary to the commonly held belief that the dinosaurs left no descendants, the seven specimens of *Archaeopteryx* (the earliest bird known) provide compelling evidence that birds (class Aves) evolved from small theropod dinosaurs. Following the principles of genealogy that are applied to humans as much as to other organisms, organisms are classified at a higher level within the groups from which they evolved. *Archaeopteryx* is therefore classified as both a dinosaur and a bird, just as humans are both primates and mammals.

The specimens of *Archaeopteryx* contain particular anatomic features that also are exclusively present in certain theropods (*Oviraptor, Velociraptor, Deinonychus,* and *Troodon,* among others). These animals share long arms and hands, a somewhat shorter, stiffened tail, a similar pelvis, and an unusual wrist joint in which the hand is allowed to flex sideways instead of up and down. This wrist motion is virtually identical to the motion used by birds (and bats) in flight, though in these small dinosaurs its initial primary function was probably in catching prey.

The origin of feathers

Beginning in the 1990s, several specimens of small theropod dinosaurs from the Early Cretaceous of Liaoning province, China, were unearthed. These fossils are remarkably well preserved, and because they include impressions of featherlike, filamentous structures that covered the body, they have shed much light on the relationship between birds and Mesozoic dinosaurs. Such structures on the skin of *Sinosauropteryx* are similar to the barbs of feathers, which suggests that feathers evolved from a much simpler structure that probably functioned as an insulator. True feathers of several types, including contour and body feathers, have been found in the 125-million-year-old feathered oviraptorid *Caudipteryx* and the apparently related *Protarchaeopteryx.* Because these animals were not birds and did not fly, it is now evident that true feathers neither evolved first in birds nor developed for the purpose of flight. Instead, feathers may have evolved for insulation, display, camouflage, species recognition, or some combination of these functions and only later became adapted for flight. In the case of *Caudipteryx,* for example, it has been established that these animals not only sat on nests but probably protected the eggs with their feathers.

Because representatives of living bird groups have long been known among the fossil species from the Paleocene and Eocene epochs (66.4 million to 36.6 million years ago), it has seemed evident that birds must have existed during the Cretaceous. Knowledge of these, based on fragments of fossil bone, has slowly come to light, and there is now a fairly definite record from Cretaceous rock strata of other ancestral birds related to the living groups of loons, grebes, flamingos, cranes, parrots, and shorebirds—and thus indication of early avian diversity. Therefore, it is clear that birds did not go through a "bottleneck" of extinction at the end of the Cretaceous that separated the archaic groups from the extant groups. Rather, the living groups were mostly present by the latest Cretaceous, and by this time the archaic groups seem to have died out.

Natural history

HABITATS

Dinosaurs lived in many kinds of terrestrial environments, and although some remains, such as footprints, indicate where dinosaurs actually lived, their bones tell us only where they died (assuming that they have not been scattered or washed far from their place of death). Not all environments are equally well preserved in the fossil record. Upland environments, forests, and plains tend to experience erosion or decomposition of organic remains, so remains from these environments are rarely preserved in the geologic record. As a result, most dinosaur fossils are known from lowland environments, usually floodplains, deltas, lake beds, stream bottoms, and even some marine environments, where their bones apparently washed in after death. Much about the environments dinosaurs lived in can be learned from studying the pollen and plant remains preserved with them and from geochemical isotopes that indicate temperature and precipitation levels. These climates, although free from the extensive ice caps of today and generally more equable, suffered extreme monsoon seasons and made much of the globe arid.

FOOD AND FEEDING

The plant eaters. From the Triassic through the Jurassic and into the Cretaceous, the Earth's vegetation changed slowly but fundamentally from forests rich in gymnosperms (cycadeoids, cycads, and conifers) to angiosperm-dominated forests of palmlike trees and magnolia-like hardwoods. Although conifers continued to flourish at high latitudes, palms were increasingly confined to subtropical and tropical regions. These forms of plant life, the vast majority of them low in calories and proteins and made largely of hard-to-digest cellulose, became the foods of changing dinosaur communities. Accordingly, certain groups of dinosaurs, such as the ornithopods, included a succession of types that were increasingly adapted for efficient food processing. At the peak of the ornithopod lineage, the hadrosaurs (duck-billed dinosaurs of the Late Cretaceous) featured large dental batteries in both the upper and lower jaws, which consisted of many tightly compressed teeth that formed a long crushing or grinding surface. The preferred food of the duckbills cannot be certified, but at least one specimen found in Wyoming offers an intriguing clue: fossil plant remains in the stomach region have been identified as pine needles.

The hadrosaurs' Late Cretaceous contemporaries, the ceratopsians (horned dinosaurs), had similar dental batteries that consisted of dozens of teeth. In this group the upper and lower batteries came together and acted as serrated shearing blades rather than crushing or grinding surfaces. Ordinarily, slicing teeth are found only in flesh-eating animals, but the bulky bodies and the unclawed, hooflike feet of dinosaurs such as *Triceratops* clearly are those of plant eaters.

The giant sauropods such as *Diplodocus* and *Apatosaurus* must have required large quantities of plant food, but there is no direct evidence as to the particular plants they preferred. Because angiosperms rich in calories and proteins did not exist during most of the Mesozoic Era, it must be assumed that these sauropods fed on the abundant conifers and palm trees. Such a cellulose-heavy diet would have required an unusual bacterial population in the intestines to break down the fibre. A digestive tract with one or more crop chambers containing stones might have aided in the food-pulverizing process, but such gastroliths, or "stomach-stones," are only rarely found in association with dinosaur skeletons. (*A Seismosaurus* specimen found with several hundred such stones is an important exception.)

The food preference of herbivorous dinosaurs can be inferred to some extent from their general body plan and from their teeth. It is probable, for example, that low-built animals such as the ankylosaurs, stegosaurs, and ceratopsians fed on low shrubbery. The tall ornithopods, especially the duckbills, and the long-necked sauropods probably browsed on high branches and treetops. No dinosaurs could have fed on grasses (family Poaceae), as these plants had not yet evolved.

The flesh eaters. The flesh-eating dinosaurs came in all shapes and sizes and account for about 40 percent of the diversity of Mesozoic dinosaurs. They must have eaten anything they could catch, because predation is a highly opportunistic lifestyle. In several instances the prey victim of a particular carnivore has been established beyond much doubt. Remains were found of the small predator *Compsognathus* containing a tiny skeleton of the lizard *Bavarisaurus* in its stomach region. In Mongolia two different dinosaur skeletons were found together, a nearly adult-size *Protoceratops* in the clutches of its predator *Velociraptor.* Two of the many skeletons of *Coelophysis* discovered at Ghost Ranch in New Mexico, U.S., contained bones of several half-grown *Coelophysis,* apparently an early Mesozoic example of cannibalism. Fossilized feces (coprolites) from a large tyrannosaur contained crushed bone of another dinosaur. Skeletons of *Deinonychus* unearthed in Montana, U.S., were mixed with fragmentary bones of a much larger victim, the herbivore *Tenontosaurus.* This last example is significant because the multiple remains of the predator *Deinonychus,* associated with the bones of a single large prey animal, *Tenontosaurus,* strongly suggest that *Deinonychus* hunted in packs.

HERDING BEHAVIOUR

It should not come as a surprise that *Deinonychus* was a social animal, because many animals today are gregarious and form groups. Fossil evidence documents similar herding behaviour in a variety of dinosaurs. The mass assemblage in Bernissart, Belgium, for example, held at least



Figure 2: Trackways of a sauropod and a carnivorous theropod at the Paluxy River site near Glen Rose, Texas.
Courtesy, Library Services Department, American Museum of Natural History, New York City; photograph, R.T. Bird (Neg. No. 324393)

three groups of *Iguanodon.* Group association and activity is also indicated by the dozens of *Coelophysis* skeletons of all ages recovered in New Mexico, U.S. The many specimens of *Allosaurus* at the Cleveland-Lloyd Quarry in Utah, U.S., may denote a herd of animals attracted to the site for the common purpose of scavenging. In the last two decades, several assemblages of ceratopsians and duckbills containing thousands of individuals have been found. Even *Tyrannosaurus rex* is now known from sites where a group has been preserved together.

Dinosaur footprints

These rare occurrences of multiple skeletal remains have repeatedly been reinforced by dinosaur footprints as evidence of herding. Trackways were first noted by Roland T. Bird in the early 1940s along the Paluxy River bed in central Texas, U.S. (Figure 2). One set of numerous wash-basin-size depressions proved to be a series of giant sauropod footsteps preserved in limestone of the Early Cretaceous Period (144 million to 97.5 million years ago). Because the tracks are nearly parallel and all progress in the same direction, Bird concluded that "all were headed toward a common objective" and suggested that the sauropod trackmakers "passed in a single herd." Large trackway sites also exist in the eastern and western United States, Canada, Australia, England, Argentina, South Africa, and China, among other places. These sites, dating from the Late Triassic Period (230 million to 208 million years ago) to the latest Cretaceous (66.4 million years ago), document herding as common behaviour among a variety of dinosaur types.

Some dinosaur trackways record hundreds, perhaps even thousands, of animals, possibly indicating mass migrations. The existence of so many trackways suggests the presence of great populations of sauropods, prosauropods, ornithopods, and probably most other kinds of dinosaurs. The majority must have been herbivores, and many of them were huge, weighing several tons or more. The impact of such large herds on the plant life of the time must have been great, suggesting constant migration in search of food.

Nesting sites discovered in the late 20th century also establish herding among dinosaurs. Nests and eggs numbering from dozens to thousands are preserved at sites that were possibly used for thousands of years by the same evolving populations of dinosaurs.

GROWTH AND LIFE SPAN

Much attention has been devoted to dinosaurs as living animals—moving, eating, growing, reproducing biological machines. But how fast did they grow? How long did they live? How did they reproduce? The evidence concerning growth and life expectancy is sparse but growing. In the 1990s histological studies of fossilized bone by Armand de Ricqlès in Paris and R.E.H. Reid in Ireland showed that dinosaur skeletons grew quite rapidly. The time required for full growth has not been quantified for most dinosaurs, but de Ricqlès and his colleagues have shown that duckbills (hadrosaurs) such as *Hypacrosaurus* and *Maiasaura* reached adult size in seven or eight years and that the giant sauropods reached nearly full size in as little as 12 years. How long dinosaurs lived after reaching adult size is difficult to determine, but it is thought that the majority of known skeletons are not fully grown, because their bone ends and arches are very often not fused; in mature individuals these features would be fused.

REPRODUCTION

The idea that dinosaurs, like most living reptiles and birds, built nests and laid eggs had been widely debated even before the 1920s, when a team of scientists led by Roy Chapman Andrews from the American Museum of Natural History, New York, made an expedition to Mongolia. Their discovery of dinosaur eggs in the Gobi Desert proved conclusively that at least one kind of dinosaur had been an egg layer and nest builder. These eggs were at first attributed to *Protoceratops,* but they are now known to have been those of *Oviraptor* (Figure 3). In 1978 John R. Horner and his field crews from Princeton University discovered dinosaur nests in western Montana. A few other finds, mostly of eggshell fragments from a number of sites, established oviparity as the only known mode of reproduction. In recent years an increasing number of dinosaur eggshells have been found and identified with the dinosaurs that laid them, and embryos have been found inside some eggs.

The almost complete absence of juvenile dinosaur remains was puzzling until the 1980s. Horner, having moved to Montana State University, demonstrated that most paleontologists simply had not been exploring the right territory. After a series of intensive searches for the remains of immature dinosaurs, he succeeded beyond all expectations. The first such bones were unearthed near Choteau, Montana, and thereafter Horner and his crews discovered hundreds of nests, eggs, and newly hatched dinosaurs (mostly duckbills). Egg Mountain, as the area was named, produced some of the most important clues to dinosauri-

Courtesy, Library Services Department, American Museum of Natural History, New York City (Neg. No. 324083)

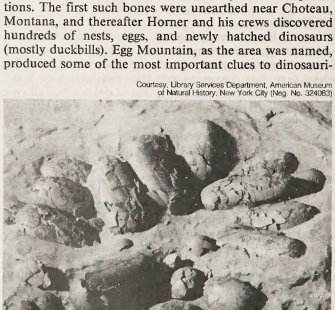
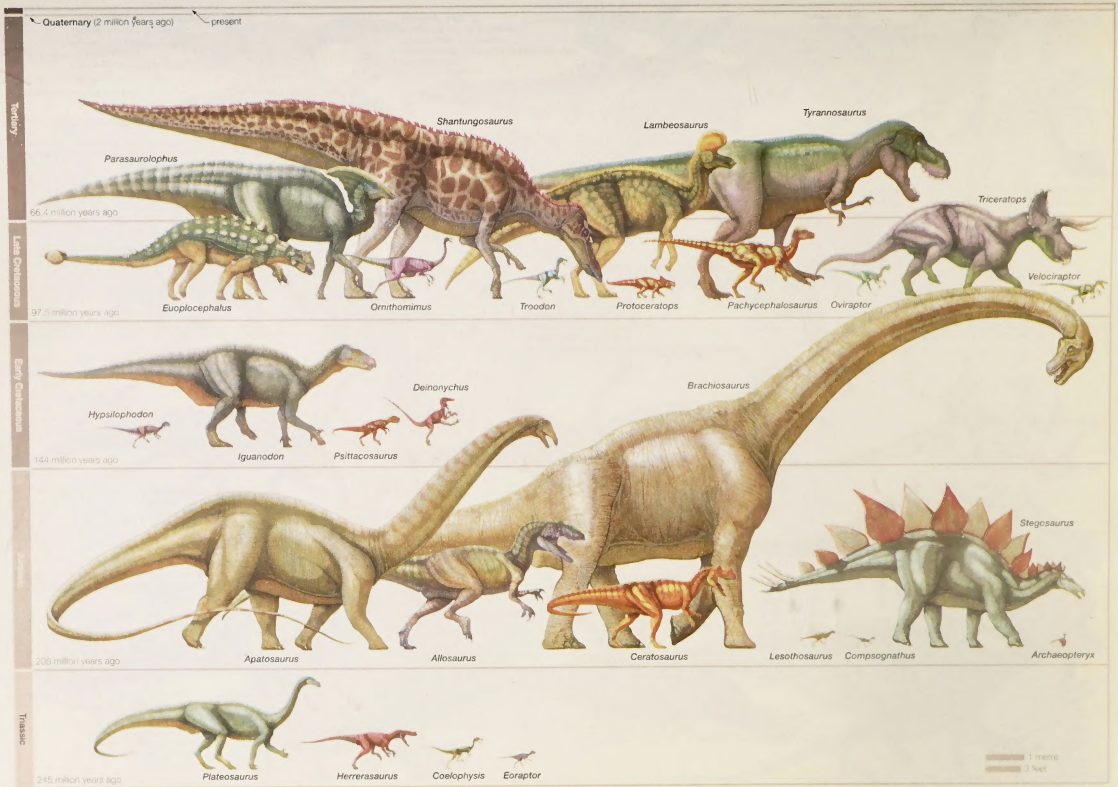
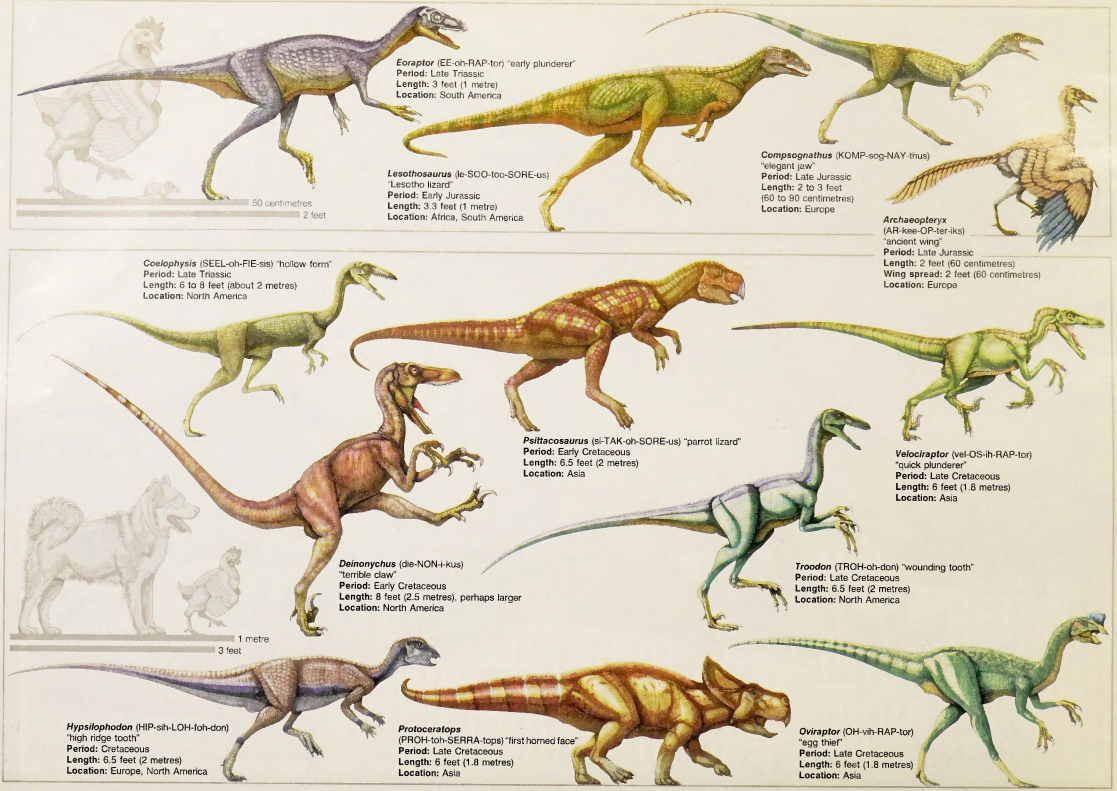


Figure 3: Nest of *Oviraptor* eggs found in Mongolia.


Quaternary (2 million years ago)
present
Tertiary
66.4 million years ago
Late Cretaceous
97.5 million years ago
Early Cretaceous
144 million years ago
Jurassic
208 million years ago
Triassic
245 million years ago
Parasaurolophus
Shantungosaurus
Lambeosaurus
Tyrannosaurus
Triceratops
Velociraptor
Euoplocephalus
Ornithomimus
Troodon
Protoceratops
Pachycephalosaurus
Oviraptor
Hypsilophodon
Iguanodon
Deinonychus
Psittacosaurus
Brachiosaurus
Stegosaurus
Apatosaurus
Allosaurus
Ceratosaurus
Lesothosaurus
Compsognathus
Archaeopteryx
Plateosaurus
Herrerasaurus
Coelophysis
Eoraptor
1 metre
3 feet



Eoraptor (EE-oh-RAP-tor) "early plunderer"
Period: Late Triassic
Length: 3 feet (1 metre)
Location: South America

Lesothosaurus (le-SOO-too-SORE-us)
"Lesotho lizard"
Period: Early Jurassic
Length: 3.3 feet (1 metre)
Location: Africa, South America

Compsognathus (KOMP-sog-NAY-thus)
"elegant jaw"
Period: Late Jurassic
Length: 2 to 3 feet
(60 to 90 centimetres)
Location: Europe

Archaeopteryx
(AR-kee-OP-ter-iks)
"ancient wing"
Period: Late Jurassic
Length: 2 feet (60 centimetres)
Wing spread: 2 feet (60 centimetres)
Location: Europe

Coelophysis (SEEL-oh-FIE-sis) "hollow form"
Period: Late Triassic
Length: 6 to 8 feet (about 2 metres)
Location: North America

Psittacosaurus (si-TAK-oh-SORE-us) "parrot lizard"
Period: Early Cretaceous
Length: 6.5 feet (2 metres)
Location: Asia

Velociraptor (vel-OS-ih-RAP-tor)
"quick plunderer"
Period: Late Cretaceous
Length: 6 feet (1.8 metres)
Location: Asia

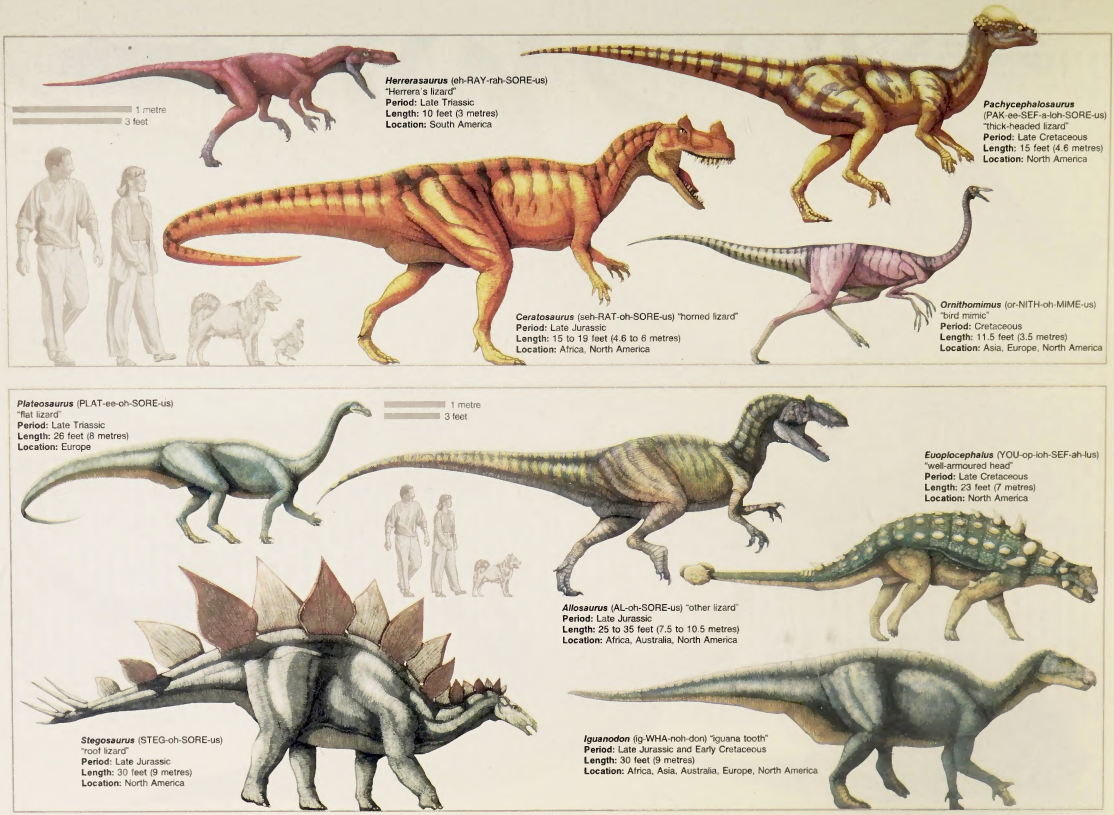
Deinonychus (die-NON-i-kus)
"terrible claw"
Period: Early Cretaceous
Length: 8 feet (2.5 metres), perhaps larger
Location: North America

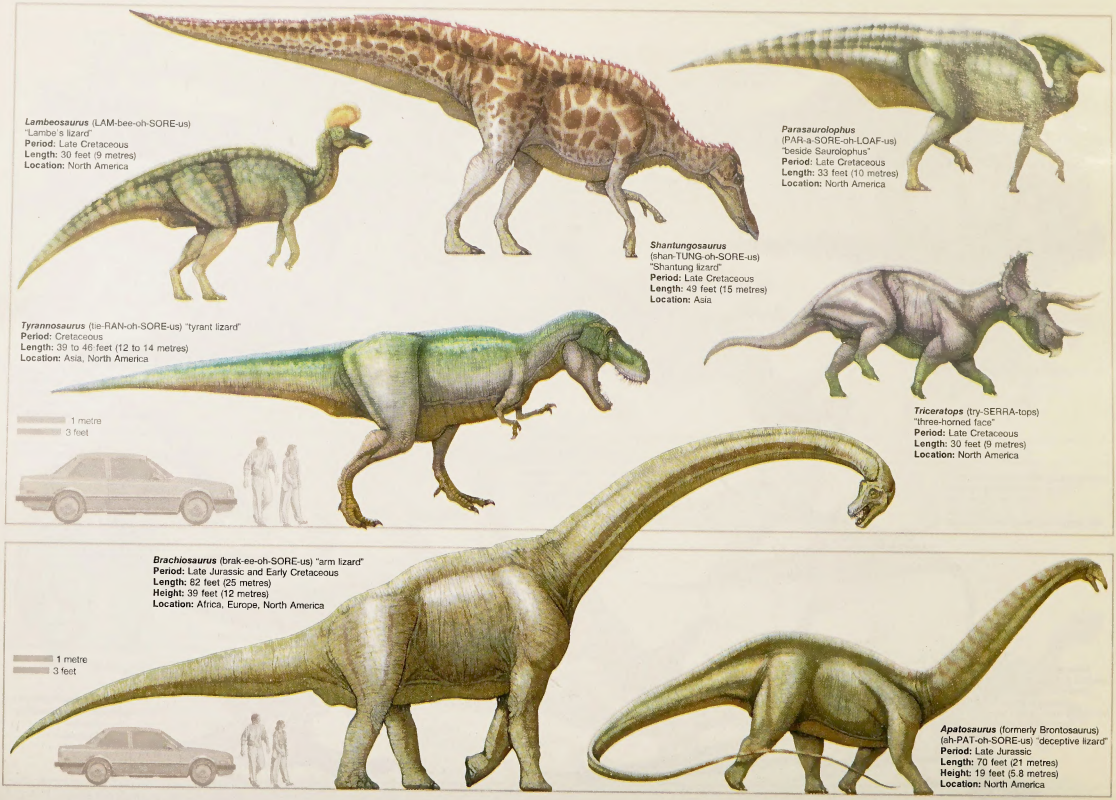
Troodon (TROH-oh-don) "wounding tooth"
Period: Late Cretaceous
Length: 6.5 feet (2 metres)
Location: North America

Hypsilophodon (HIP-sih-LOH-foh-don)
"high ridge tooth"
Period: Cretaceous
Length: 6.5 feet (2 metres)
Location: Europe, North America

Protoceratops
(PROH-toh-SERRA-tops) "first horned face"
Period: Late Cretaceous
Length: 6 feet (1.8 metres)
Location: Asia

Oviraptor (OH-vih-RAP-tor)
"egg thief"
Period: Late Cretaceous
Length: 6 feet (1.8 metres)
Location: Asia


1 metre
3 feet
Herrerasaurus (eh-RAY-rah-SORE-us)
"Herrera's lizard"
Period: Late Triassic
Length: 10 feet (3 metres)
Location: South America
Pachycephalosaurus
(PAK-ee-SEF-a-loh-SORE-us)
"thick-headed lizard"
Period: Late Cretaceous
Length: 15 feet (4.6 metres)
Location: North America
Ceratosaurus (seh-RAT-oh-SORE-us) "horned lizard"
Period: Late Jurassic
Length: 15 to 19 feet (4.6 to 6 metres)
Location: Africa, North America
Ornithomimus (or-NITH-oh-MIME-us)
"bird mimic"
Period: Cretaceous
Length: 11.5 feet (3.5 metres)
Location: Asia, Europe, North America
Plateosaurus (PLAT-ee-oh-SORE-us)
"flat lizard"
Period: Late Triassic
Length: 26 feet (8 metres)
Location: Europe
1 metre
3 feet
Euoplocephalus (YOU-op-loh-SEF-ah-lus)
"well-armoured head"
Period: Late Cretaceous
Length: 23 feet (7 metres)
Location: North America
Allosaurus (AL-oh-SORE-us) "other lizard"
Period: Late Jurassic
Length: 25 to 35 feet (7.5 to 10.5 metres)
Location: Africa, Australia, North America
Stegosaurus (STEG-oh-SORE-us)
"roof lizard"
Period: Late Jurassic
Length: 30 feet (9 metres)
Location: North America
Iguanodon (ig-WHA-noh-don) "iguana tooth"
Period: Late Jurassic and Early Cretaceous
Length: 30 feet (9 metres)
Location: Africa, Asia, Australia, Europe, North America



Lambeosaurus (LAM-bee-oh-SORE-us)
"Lambe's lizard"
Period: Late Cretaceous
Length: 30 feet (9 metres)
Location: North America

Parasaurolophus
(PAR-a-SORE-oh-LOAF-us)
"beside Saurolophus"
Period: Late Cretaceous
Length: 33 feet (10 metres)
Location: North America

Shantungosaurus
(shan-TUNG-oh-SORE-us)
"Shantung lizard"
Period: Late Cretaceous
Length: 49 feet (15 metres)
Location: Asia

Tyrannosaurus (tie-RAN-oh-SORE-us) "tyrant lizard"
Period: Cretaceous
Length: 39 to 46 feet (12 to 14 metres)
Location: Asia, North America

1 metre
3 feet

Triceratops (try-SERRA-tops)
"three-horned face"
Period: Late Cretaceous
Length: 30 feet (9 metres)
Location: North America

Brachiosaurus (brak-ee-oh-SORE-us) "arm lizard"
Period: Late Jurassic and Early Cretaceous
Length: 82 feet (25 metres)
Height: 39 feet (12 metres)
Location: Africa, Europe, North America

1 metre
3 feet

Apatosaurus (formerly Brontosaurus)
(ah-PAT-oh-SORE-us) "deceptive lizard"
Period: Late Jurassic
Length: 70 feet (21 metres)
Height: 19 feet (5.8 metres)
Location: North America

an habits yet found. For example, the sites show that a number of different dinosaur species made annual treks to this same nesting ground (though perhaps not all at the same time). Because of the succession of similar nests and eggs lying one on top of the other, it is thought that particular species returned to the same site year after year to lay their clutches. As Horner concluded, "site fidelity" was an instinctive part of dinosaurian reproductive strategy. This was confirmed more recently with the discovery of sauropod nests and eggs spread over many square kilometres in Patagonia, Argentina.

BODY TEMPERATURE

Beyond eating, digestion, assimilation, reproduction, and nesting, many other processes and activities went into making the dinosaur a successful biological machine. Breathing, fluid balance, temperature regulation, and other such capabilities are also required. Dinosaurian body temperature regulation, or lack thereof, has been a hotly debated topic among students of dinosaur biology. Because it is obviously not possible to take an extinct dinosaur's temperature, all aspects of their metabolism and thermophysiology can be assessed only indirectly.

Ectothermy and endothermy. All animals thermoregulate. The internal environment of the body is under the influence of both external and internal conditions. Land animals thermoregulate in several ways. They do so behaviorally, by moving to a colder or warmer place, by exercising to generate body heat, or by panting or sweating to lose it. They also thermoregulate physiologically, by activating internal metabolic processes that warm or cool the blood. But these efforts have limits, and, as a result, external temperatures and climatic conditions are among the most important factors controlling the geographic distribution of animals.

Present-day metabolisms

Today's so-called warm-blooded animals are the mammals and birds; reptiles, amphibians, and most fishes are called cold-blooded. These two terms, however, are imprecise and misleading. Some "cold-blooded" lizards have higher normal body temperatures than do some mammals, for instance. Another pair of terms, "ectothermy" and "endothermy," describes whether most of an animal's heat is absorbed from the environment ("ecto-") or generated by internal processes ("endo-"). A third pair of terms, "poikilothermy" and "homeothermy," describes whether the body temperature tends to vary with that of the immediate environment or remains relatively constant.

Today's mammals and birds have a high metabolism and are considered endotherms, which produce body heat internally. They possess biological temperature sensors that control heat production and switch on heat-loss mechanisms such as perspiration. Today's reptiles and amphibians, on the other hand, are ectotherms that mostly gain heat energy from sunlight, a heated rock surface, or some other external source. The endothermic state is effective but metabolically expensive, as the body must produce heat continuously, which requires correspondingly high quantities of fuel in the form of food. On the other hand, endotherms can be more active and survive lower external temperatures. Ectotherms do not require as much fuel, but most cannot deal as well with cold surroundings.

From the time of the earliest discoveries in the 19th century, dinosaur remains were classified as reptilian because their anatomic features are typical of living reptiles such as turtles, crocodiles, and lizards. Because dinosaurs all have lower jaws constructed of several bones, a reptilian jaw joint, and a number of other nonmammalian, nonbirdlike characteristics, it was assumed that living dinosaurs were similar to living reptiles—scaly, cold-blooded, ectothermic egg layers (predominantly), not furry, warm-blooded live-bearers. A chauvinistic attitude seems to prevail that the warm-bloodedness of mammals is better than the cold-blooded reptilian state, even though turtles, snakes, and other reptiles do very well regulating their body temperature in a different way. Moreover, both birds and mammals evolved from ectothermic, poikilothermic ancestors. At what point did metabolism heat up?

Clues to dinosaurian metabolism. The question of whether any extinct dinosaur was a true endotherm or homeotherm cannot be answered, but some interesting anatomic facts suggest these "warmer" possibilities. Probably the most direct evidence of dinosaurian physiology comes from bones themselves, particularly in regard to how they grew. The long bones (such as arm and leg bones) of most dinosaurs are composed almost exclusively of a well-vascularized type of bone matrix (fibro-lamellar) also found in most mammals and large birds. This type of bone tissue always indicates rapid growth, and it is very different from the more compact, poorly vascularized, parallel-fibred bone found in crocodiles and other reptiles and amphibians. It is generally thought that well-vascularized, rapidly growing bone can be sustained only by high metabolic rates that bring a continual source of nutrients and minerals to the growing tissues. It is difficult to explain these histological features in any other metabolic terms. On the other hand, most dinosaurs retain lines of arrested growth (LAGs) in most of their long bones. LAGs are found in other reptiles, amphibians, and fishes, and they often reflect a seasonal period during which metabolism slows, usually because of environmental stresses. This slowdown produces "rest lines" as LAGs in the bones. The presence of these lines in dinosaur bones has been taken as an indication that they were metabolically incapable of growing throughout the year. However, LAGs in dinosaurs are less pronounced than in other reptiles; LAGs can also appear in different numbers in different bones of the same skeleton, and they are sometimes even completely absent. Finally, some living birds and mammals, which are clearly endotherms, have LAGs very much like those of dinosaurs, so LAGs are probably not strong indicators of metabolism in any of these animals.

Fossil evidence of metabolism

Other, less direct lines of evidence may reveal other clues about dinosaurian metabolism. Two dinosaurian groups, the hadrosaurs and the ceratopsians, had highly specialized sets of teeth that were obviously effective at processing food. Both groups were herbivorous, but unlike living reptiles they chopped and ground foliage thoroughly. Such highly efficient dentitions may suggest a highly effective digestive process that would allow more energy to be extracted from the food. This feature by itself, however, may not be crucial. Pandas, for example, are not very efficient in digesting plant material, but they survive quite well on a diet of almost nothing but bamboo.

Another line of evidence is that dinosaurs had anatomic features reflecting a high capacity for activity. The first dinosaurs walked upright, holding their legs under their bodies; they could not sprawl. This indicates that, by standing and walking all day, they probably expended more energy than reptiles, which typically sit and wait for prey. As some lineages of dinosaurs grew larger, they reverted to four-legged (quadrupedal) locomotion, but their stance was still upright. They also put one foot directly in front of the other when they walked (parasagittal gait), instead of swinging the limbs to the side. Such posture and gait are present in all nonaquatic endotherms (mammals and birds) today, whereas a sprawling or semierect posture is typical of all ectotherms (reptiles and amphibians). Bipedal stance and parasagittal gait are not sustained in any living ectotherm, perhaps because they require a relatively higher level of sustained energy.

Circumstantial evidence of metabolism

The high speeds at which some dinosaurs must have traveled have also been invoked as evidence of high metabolic levels. For example, the ostrichlike dinosaurs, such as *Struthiomimus, Ornithomimus, Gallimimus,* and *Dromiceiomimus,* had long hind legs and must have been very fleet. The dromaeosaurs, such as *Deinonychus, Velociraptor,* and *Dromaeosaurus,* also were obligatory bipeds. They killed prey with talons on their feet, and one can argue that it must have taken a high level of metabolism to generate the degree of activity and agility required of such a skill. However, most ectotherms can move very rapidly in bursts of activity such as running and fighting, so this feature may not provide conclusive evidence either.

Related to the upright posture of many dinosaurs is the fact that the head was often positioned well above the level of the heart. In some sauropods (*Apatosaurus, Diplodocus, Brachiosaurus,* and *Barosaurus,* for instance), the brain must have been several metres above the heart. The phys-

iological importance of this is that a four-chambered heart would be required for pumping freshly oxygenated blood to the brain. Brain death follows very quickly when nerve cells are deprived of oxygen, and to prevent it most dinosaurs must have required two ventricles. In a four-chambered heart, one ventricle pumps oxygen-poor venous blood at low pressure to the lungs to absorb fresh oxygen (high pressure would rupture capillaries of the lungs). A powerful second ventricle pumps freshly oxygenated blood to all other parts of the body at high pressure. To overcome the weight of the column of blood that must be moved from the heart to the elevated brain, high pressure is certainly needed. In short, like birds and mammals, many dinosaurs apparently had the required four-chambered heart necessary for an animal with a high metabolism.

The significance of thermoregulation can be seen by comparing today's reptiles with mammals. The rate of metabolism is usually measured in terms of oxygen consumed per unit of body weight per unit of time. The resting metabolic rate for most mammals is about 10 times that of modern reptiles, and the range of metabolic rates among living mammals is about double that seen among reptiles. These differences mean that endothermic mammals have much more endurance than their cold-blooded counterparts. Some dinosaurs may have been so endowed, and although they seem to have possessed the cardiovascular system necessary for endothermy, that capacity does not conclusively prove that they were endothermic. There exists the possibility that dinosaurs were neither complete ectotherms nor complete endotherms. Rather, they may have evolved a range of metabolic strategies, much as mammals have (as is illustrated by the differences between sloths and cheetahs, bats and whales, for example).

Classification

The chief difference between the two major groups of dinosaurs is in the configuration of the pelvis. It was primarily on this distinction that the English biologist H.G. Seeley established the two dinosaurian orders and named them Saurischia ("lizard hips") and Ornithischia ("bird hips") in 1887; this differentiation is still maintained (Figure 4).

Pelvic structure of the two groups

As in all four-legged animals, the dinosaurian pelvis was a paired structure consisting of three separate bones on each side that attached to the sacrum of the backbone. The ilium was attached to the spine, and the pubis and ischium were below, forming a robust bony plate. At the centre of each plate was a deep cup—the hip socket (acetabulum). The hip socket faced outward and was open at its centre for the articulation of the thighbone. The combined saurischian pelvic bones presented a triangular outline as seen from the side, with the pubis extending down and forward and the ischium projecting down and backward from the hip socket. The massive ilium formed a deep vertical plate of bone to which the muscles of the pelvis, hind leg, and tail were attached. The pubis had a stout shaft, commonly terminating in a pronounced expansion or bootlike structure (presumably for muscle attachment) that solidly joined its opposite mate. The ischium was slightly less robust than the pubis, but it too joined its mate along a midline. There were minor variations in this structure between the various saurischians.

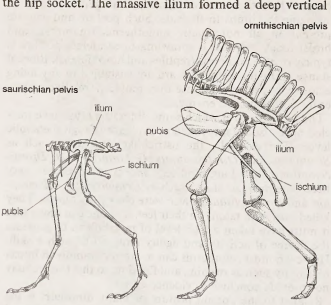


Figure 4: Dinosaur pelvic structures.
Encyclopædia Britannica

The ornithischian pelvis was constructed of the same three bones on each side of the sacral vertebrae, to which they were attached. The lateral profile of the pelvis was quite different from that of the saurischians, which had a long but low iliac blade above the hip socket and a modified ischium-pubis structure below. Here the long, thin ischium extended backward and slightly downward from the hip socket. In the most primitive, or basal, ornithischians, the pubis had a moderately long anterior blade, but this was reduced in later ornithischians. Posteriorly it stretched out into a long, thin postpubic process lying beneath and closely parallel to the ischium. The resulting configuration superficially resembled that of birds, whose pubis is a thin process extending backward beneath the larger ischium. These anatomic dissimilarities are thought to reflect important differences in muscle arrangements in the hips and hind legs of these two orders. However, the soft parts of these dinosaurs are not well enough understood to reveal any functional or physiological basis for the differences. Other marked dissimilarities between saurischians and ornithischians are found in their jaws and teeth, their limbs, and especially their skulls. Details regarding these differences are given in the following discussions of the major dinosaur groups.

Dinosaur subgroups

- Saurischia
  - Sauropodomorpha
    - Prosauropoda
    - Sauropoda
  - Theropoda
    - Ceratosauria
    - Tetanurae
- Ornithischia
  - Cerapoda
    - Ornithopoda
    - Pachycephalosauria
    - Ceratopsia
  - Thyreophora
    - Stegosauria
    - Ankylosauria

The table shows how the major dinosaur groups are subdivided. The table and classification (Figure 5) are based on the groups' relationships to each other, as far as they are known. Fossil remains are often difficult to interpret, especially when only a few fragmentary specimens of a type have been found. No universally accepted classification of dinosaurs exists. Occasionally, for example, the Sauropodomorpha have been divided into more or fewer lower-rank categories (*e.g.*, families, subfamilies), and the suborder Theropoda has been divided into two infraorders, the Carnosauria and the Coelurosauria. Increasingly, taxonomists have abandoned the traditional Linnaean ranks of family, order, and so on because they are cumbersome and not comparable among different kinds of organisms. Instead, the names of the groups alone are used without denoting a category. Generally, a phylogeny such as the accompanying diagram clearly shows which groups are subsumed under others. Additionally, words with similar roots but different endings may indicate more or less inclusive groups. Ornithomimosauria, for example, denotes a more inclusive group than Ornithomimidae. Because the results of different phylogenetic analyses vary among researchers, and will continue to change as new specimens and taxa are discovered, the classification can be expected to change accordingly. This is a normal part of scientific activity and reflects continuing growth of knowledge and reappraisal of current understanding.

SAURISCHIA

Saurischians are known from specimens ranging from the Late Triassic to the present day, because, as will be seen,

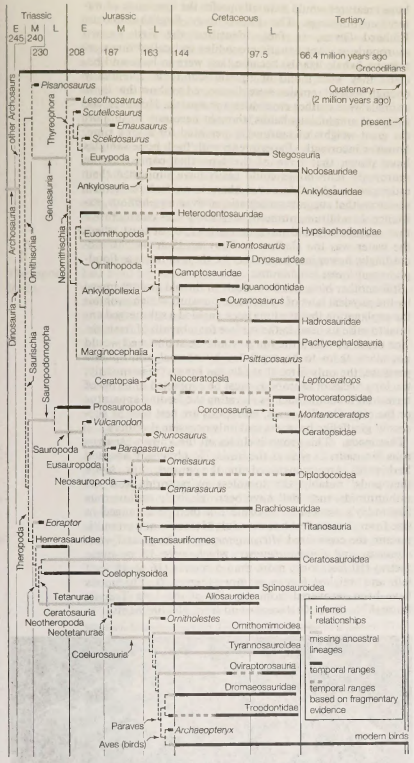


Figure 5: Dinosaur phylogeny, or "family tree."
Courtesy of Paul C. Sereno (2001), University of Chicago

birds are highly derived saurischian dinosaurs. Two distinctly different groups are traditionally included in the saurischians—the Sauropodomorpha (herbivorous sauropods and prosauropods) and the Theropoda (carnivorous dinosaurs). These groups are placed together on the basis of a suite of features that they share uniquely. These include elongated posterior neck vertebrae, accessory articulations on the trunk vertebrae, and a hand that is nearly half as long as the rest of the arm (or longer).

Sauropodomorpha. Included in this group are the well-known sauropods, or "brontosaur" types, and their probable ancestral group, the prosauropods. All were plant eaters, though their relationship to theropods, along with the fact that the closest relatives of dinosaurs were evidently carnivorous, suggests that they evolved from meat eaters. Sauropodomorpha are distinguished by leaf-shaped tooth crowns, a small head, and a neck that is at least as long as the trunk of the body and longer than the limbs.

Prosauropoda. Most generalized of the Sauropodomorpha were the so-called prosauropods. Found from the Late Triassic to Early Jurassic periods, their remains are probably the most ubiquitous of all Triassic dinosaurs. The best-known examples include *Plateosaurus* of Germany and *Massospondylus* of South Africa. Prosauropods were not especially large; they ranged from less than 2 metres (7 feet) in length up to about 8 metres (26 feet) and up to several tons in maximum weight. Many of these animals are known from very complete skeletons (especially the smaller, more lightly built forms). Because their forelimbs are conspicuously shorter than their hind limbs, they have often been reconstructed poised on their hind legs in a bipedal stance. Their anatomy, however, clearly indicates that some of them could assume a quadrupedal (four-footed) position. Footprints generally attributed to prosauropods appear to substantiate both forms of locomotion.

Possible sauropod ancestry

Prosauropods have long been seen as including the first direct ancestors of the giant sauropods, probably among the melanorosaurids. That view has long prevailed largely because of their distinctly primitive sauropod-like appearance and also because of their Late Triassic–Early Jurassic occurrence. No better candidate has been discovered, and the first true sauropods are not found until the Early Jurassic, so the transition between prosauropods and sauropods has been generally accepted. In the 1990s, however, several studies have suggested that prosauropods may be a distinct group that shared common ancestors with sauropods earlier in the Triassic. If this view is correct, it is mystifying why the smaller prosauropods are so widespread throughout the Late Triassic, yet none of the larger and more conspicuous sauropods have been found from that period.

In general body form, prosauropods were mostly rather stocky, with a long, moderately flexible neck containing surprisingly long and flexible cervical ribs. The head was small in comparison with the body. The jaw was long and contained rows of thin, leaflike teeth suited for chopping up (but not grinding or crushing) plant tissues, although there is an indication of direct tooth-on-tooth occlusion.

Prosauropod forelimbs were stout, with five complete digits. The hind limbs were about 50 percent longer than the forelimbs and even more heavily built. The foot was of primitive design, and its five-toed configuration could be interpreted as a forerunner of the sauropod foot. Walking apparently was done partly on the toes (semidigitigrade), with the metatarsus held well off the ground. The vertebral column was unspecialized and bore little indication of the cavernous excavations that were to come in later sauropod vertebrae, nor did it show projections that were to buttress the sauropod vertebral column. The long tail probably served as a counterweight or stabilizer whenever the animal assumed a bipedal position.

Sauropoda. The more widely known sauropods—the huge "brontosaurs" and their relatives—varied in length from 6 or 7 metres (about 20 feet) in the primitive ancestral sauropod *Vulcanodon* of Africa, *Barapasaurus* of India, and *Ohmdenosaurus* of Germany up to 28 to 30 metres (90 to 100 feet) or more in Late Jurassic North American forms such as *Apatosaurus* (formerly known as *Brontosaurus*), *Diplodocus*, *Seismosaurus*, and *Sauroposeidon.* Weights ranged from about 20 tons or less in *Barapasaurus* to 80 tons or more for the gigantic *Brachiosaurus* of Africa and North America (Figure 6). Sauropods were worldwide in distribution but have not as yet been found in Antarctica. In geologic time they ranged from the Late Triassic *Riojasaurus* to the Late Cretaceous *Alamosaurus* of North America and *Laplatasaurus* of South America. Their greatest diversity and abundance took place 120 million–150 million years ago, during the Late Jurassic and Early Cretaceous periods.

Sauropods are notable for their body form as well as their enormous size. Their large bodies were heart-shaped in cross section, like elephants, with long (sometimes extremely long) necks and tails. Their columnar legs, again like those of elephants, had little freedom to bend at the knee and elbow. The legs were maintained in a nearly vertical position beneath the shoulder and hip sockets. Because of their great bulk, sauropods unquestionably were obligate quadrupeds.

The sauropod limb bones were heavy and solid. The feet were broad, close to plantigrade (adapted for walking on

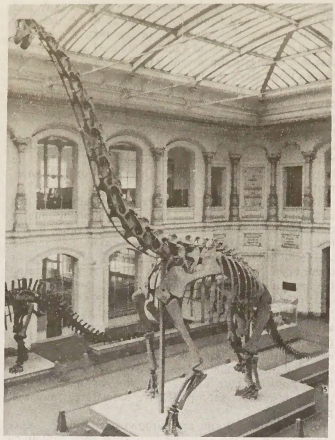


Figure 6: *Brachiosaurus* skeleton.
© Museum fur Naturkunde an der Humboldt-Universitat zu Berlin

the soles), and graviportal (adapted for bearing great weight). The toes were generally short, blunt, and broad, but some sauropods had a large straight claw on the first digit of the forefoot and the first and second toes of the hind foot. These animals must have moved relatively slowly and with only short steps because of the comparative inflexibility of the limbs. Running must have been stiff-legged at no better than an elephantine pace of 16 km (10 miles) per hour, if that. Their tremendous bulk placed them out of the reach of predators and eliminated any need for speed. Evidently their fast growth was adaptive to predator avoidance.

Sauropod backbone and tail

The vertebrae of the backbone were highly modified, with numerous excavations and struts to reduce bone weight. Complex spines and projections for muscle and ligament attachment compensated for any loss of skeletal strength that resulted from reductions in bone density and mass. The long and sometimes massive tail, characteristic of so many sauropods, would appear to have been carried well off the ground. Tail drag marks associated with sauropod trackways are not known, and damaged (stepped-on) tails are also not known, even though these animals apparently traveled in herds (albeit of undetermined density). Another possible use of the tail, like the neck, may have been thermal regulation, as improved heat loss through its large surface area could have been a result. The tail was also the critical anchor of the large, powerful hind leg muscles that produced most of the walking force required for moving the many tons of sauropod weight. The muscle arrangement of the tail was precisely that of modern alligators and lizards.

The most important part of any skeleton is the skull because it provides the most information about an animal's mode of life and general biology. Sauropod skulls were of several main types, including the high, boxy *Camarasaurus* type (often incorrectly associated with *Apatosaurus*); the shoe-shaped *Brachiosaurus* type, with its large, delicately arched nasal bones; and the low, narrow, streamlined, almost horselike *Diplodocus* type. The first had broad, spatulate teeth, while the latter two had narrow, pencil-shaped teeth largely confined to the front parts of the jaws, especially in diplodocids.

Until recently, sauropods were visualized as swamp or lake dwellers because their legs were thought to be incapable of supporting their great weights or because such huge creatures would naturally prefer the buoyancy of watery surroundings. The 19th-century English biologist Richard Owen, in fact, identified the first known sauropods as giant aquatic crocodiles and called them cetiosaurs (whale lizards) because they were so large and because they were found in aquatic sediments. Eventually enough skeletal remains were discovered to show that these animals were neither crocodiles nor aquatic. However, the image of amphibious habits, thought necessary to support the great weights of sauropods, persisted for a long time, however incorrectly. Experiments with fresh bone samples have shown that bone of the type that composed the sauropods' limb bones could easily have supported their estimated weights. Moreover, there is no feature in their skeletons that suggests an aquatic, or even amphibious, existence. In addition, numerous trackway sites clearly prove that sauropods could navigate on land, or at least where the water was too shallow to buoy up their weight. Accordingly, newer interpretations see these animals as floodplain and forest inhabitants.

Still another blow has been dealt to the old swamp image by the physical laws of hydrostatic pressure, which prohibit the explanation that the long neck enabled a submerged animal to raise its head to the surface for a breath of fresh air. The depth at which the lungs would be submerged would not allow them to be expanded by normal atmospheric pressure, the only force that fills the lungs. Consequently, the long necks of sauropods must be explained in terms of terrestrial functions such as elevating the feeding apparatus or the eyes. On all counts, sauropods are best seen as successful giraffelike browsers and only occasional waders.

Theropoda. This group includes all the known carnivorous dinosaurs as well as the birds. No obviously adapted herbivores are recognized in the group, but some theropods, notably the toothless oviraptorids and ornithomimids, may well have been relatively omnivorous like today's ostriches. Mesozoic Era theropods ranged in size from the smallest known adult Mesozoic nonavian dinosaur, the crow-sized *Microraptor,* up to the great *Tyrannosaurus* and *Giganotosaurus,* which were 15 or more metres (50 feet) long, more than 5 metres (16 to 18 feet) tall, and weighed 6 tons or more (Figure 7). Theropods have been recovered from deposits of the Late Triassic through the latest Cretaceous and from all continents.

Courtesy of the Field Museum of Natural History, Chicago; photograph, John Weinstein

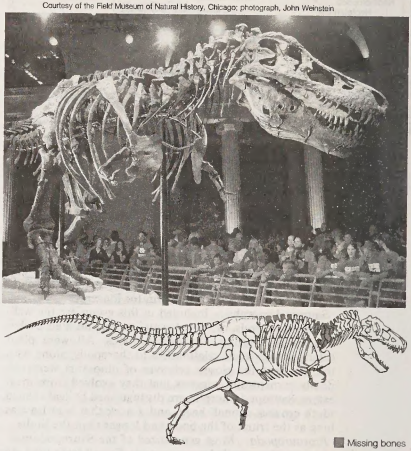


Figure 7: (Top) "Sue," a *Tyrannosaurus rex* skeleton found in South Dakota, U.S. (Bottom) Diagram of Sue's skeleton, indicating missing bones.

Theropods may be defined as birds and all saurischians more closely related to birds than to sauropods. They have a carnivorous dentition and large, recurved claws on the fingers. They also share many other characteristics, such as a distinctive joint in the lower jaw, epipophyses on the neck vertebrae, and a unique "transition point" in the tail where the vertebrae become longer and more lightly built. Other similarities include the reduction or loss of the outer two fingers, long end joints of the fingers, and a straplike fibula attached to a crest on the side of the tibia.

Herrerasaurus and several fragmentary taxa from South America, including *Staurikosaurus* and *Ischisaurus,* from the Middle to Late Triassic of Argentina are carnivores that have often been included in the Dinosauria, specifically in Theropoda. Whereas these animals closely resemble dinosaurs and have many carnivorous features, they also lack a number of features present in dinosaurs, saurischians, and theropods.

Theropod limbs

In stance and gait, theropods were obligatory bipeds. All had hind leg bones that were hollow to varying degrees—extremely hollow and lightly built in small to medium-size members (*Compsognathus, Coelurus,* and *Ornitholestes,* among others) and more solid in the larger forms (such as *Allosaurus, Daspletosaurus,* and *Tarbosaurus*). Their bodies conformed to a common shape in which the hind legs were dominant and designed for support and locomotion. The forelimbs, on the other hand, had been modified from the primitive design and entirely divested of the functions of locomotion and body support. Hind limbs were either very robust and of graviportal (weight-bearing) proportions, as in *Allosaurus, Megalosaurus,* and the tyrannosaurids, or very slender, elongated, and of cursorial (adapted for running) proportions, as in *Coelurus, Coelophysis, Ornitholestes,* and the ornithomimids. Theropod feet, despite the group's name, which means "beast (*i.e.,* mammal) foot," usually looked much like those of birds, which is not surprising, because birds inherited their foot structure from these dinosaurs. Three main toes were directed forward and splayed in a V-shaped arrangement; an additional inside toe was directed medially or backward. The whole foot was supported by the toes (digitigrade), with the "heel" elevated well above the ground. Toes usually bore sharp, somewhat curved claws.

The forelimbs varied widely from the slender, elongated ones of *Struthiomimus,* for example, to shorter, more massively constructed grasping appendages like those of *Allosaurus,* to the greatly abbreviated arms and hands of *Tyrannosaurus,* to the abbreviated, stout limb and single finger of *Mononykus,* to the range of wings now seen in birds. The hands typically featured long, flexible fingers with pronounced, often strongly curved claws, which bore sharp piercing talons. Early theropods such as *Coelophysis* had four fingers, with the fifth reduced to a nubbin of the metacarpal and the fourth greatly reduced. Most theropods were three-fingered, having lost all remnants of the fourth and fifth fingers. Tyrannosaurids (including *Albertosaurus, Daspletosaurus, Tarbosaurus,* and *Tyrannosaurus*) were notable for their two-fingered hands and unusually short arms; they had lost the third finger. The odd *Mononykus* lost even its second finger, retaining only a bizarre thumb. This separation of function between fore and hind limbs was a feature of the first dinosaurs. Although the first theropods, sauropodomorphs, and ornithischians were all bipedal, only theropods remained exclusively so.

The jaws of theropods are noted for their complement of sharp, bladelike teeth. In nearly all theropods these laterally compressed blades had serrations along the rear edge and often along the front edge as well. Tyrannosaur teeth differed in having a rounder, less-compressed cross section, better adapted to puncture flesh and tear it from bone. Troodontid teeth had recurved serrations slightly larger than those typical of theropods. *Archaeopteryx* and other basal birds had narrow-waisted teeth with greatly reduced serrations or none at all. Some theropods, such as most ornithomimids and oviraptorids, had lost most or all of their teeth.

In recent years a series of unusually well-preserved theropod dinosaurs have been discovered in deposits from the Early Cretaceous Period (144 million to 97.5 million years ago) in Liaoning province, China. These theropods have filamentous integumentary structures of several kinds that resemble feathers. Such structures indicate that today's birds very likely evolved from theropod dinosaurs. See above *Dinosaur descendants.*

Ceratosauria. Ceratosauria includes *Ceratosaurus* and all theropods more closely related to it than to birds. This group includes basal theropods such as *Coelophysis* and *Dilophosaurus.* It may also include the abelisaurids of South America and elsewhere, but this is not certain. Originally thought to be a natural group, Ceratosauria, as traditionally constituted, may represent a more general grouping of basal theropods, including the ancestral stock of most later theropods. The Late Triassic *Coelophysis,* about 1.5 meters long, is generally regarded as an archetypal primitive theropod. It had a long neck and a long, low head with numerous small, sharp, recurved teeth. The legs were long, the arms relatively short, and the tail very long. *Dilophosaurus,* from the Early Jurassic Period (208 million to 187 million years ago), is considerably larger (about 4 metres total length) and is distinguished by a pair of thin bony crests running along the top of the skull. Because no other theropod had such structures, these were apparently not necessary for any physiological function and so are thought to have been for display or species recognition. There is no evidence that *Dilophosaurus* spat venom.

Tetanurae. These comprise birds and all the theropods closer to birds than to *Ceratosaurus.* They would include the true carnosaurs and coelurosaurs described below as well as a few relatively large carnivorous basal forms (such as *Torvosaurus, Spinosaurus, Baryonyx, Afrovenator,* and *Megalosaurus*). The tetanuran theropods are distinguished by several features, including the complete loss of digits four and five of the hand, an upper tooth row extending backward only to the eye, and a fibula that is reduced and clasped by the tibia. The name Tetanurae, or "stiff tails," refers to another unusual feature, a transition point in the tail sequence where the vertebrae change form in a distinctive way.

Carnosaurs

Carnosauria includes *Allosaurus* and all theropods more closely related to it than to birds, including forms such as *Acrocanthosaurus, Sinraptor,* and *Giganotosaurus.* The first known members appear in the Late Jurassic and persist into the Cretaceous. Originally, this group was designed to include all the big predatory dinosaurs, but it was recently recognized that only size, not their relationships, was the trait unifying this group. Some, such as *Dilophosaurus* and *Carnotaurus,* were probably more closely related to basal ceratosaurs. Others, such as *Baryonyx* and *Spinosaurus,* represented an unusual diversification of fish-eating forms that were almost crocodilian in some of their habits. Still others, such as *Tyrannosaurus* and its relatives, the albertosaurs and daspletosaurs, were probably just giant coelurosaurs, as had been hypothesized by German paleontologist Friedrich von Huene early in the 20th century. As these groups were removed from the original Carnosauria, only *Allosaurus* and its relatives of the great Late Jurassic and Early Cretaceous diversification were left. Along with *Torvosaurus* and the megalosaurs, they must have been among the most deadly and rapacious large predators of their time. They are distinguished by relatively few characteristics. It is commonly thought that carnosaurs had very short limbs, but this is not particularly true—they were proportionally much shorter in tyrannosaurs, which are no longer considered carnosaurs. True carnosaurs had limbs comparable in size to those of more basal theropods. Sauropod vertebrae have been found with carnosaur tooth marks in them, which attests to the predatory habits of these dinosaurs.

The coelurosaurs ("hollow-tailed reptiles") include generally small to medium-size theropods, though the recent inclusion of tyrannosaurs would seem to discount this generalization. Coelurosauria is defined as birds and all tetanurans more closely related to birds than to the carnosaurs. The first known members, including birds, appear in the Late Jurassic; the great Cretaceous diversification of the other coelurosaurs ended with the Cretaceous extinctions.

In coelurosaurs the pelvis is modified so that the ischium is reduced to two-thirds or less the size of the pubis; the

eyes are larger, and no more than 15 tail vertebrae bear transverse projections. Each of the various coelurosaurian groups has very distinct features that set it apart from the others. The most basal known form, the Late Jurassic *Compsognathus,* was the size of a chicken and contemporaneous with the first known bird, *Archaeopteryx.* However, the two animals were not as closely related as some other coelurosaurs were to birds.

Tyrannosaurs and the related albertosaurs were the largest of the Late Cretaceous theropods of the northern continents. They are distinguished by an exceptionally large, high skull and teeth with a much more rounded cross section than the typical daggerlike teeth of other theropods. Their forelimbs are very short, and the third finger is reduced to a splint or lost entirely. Tyrannosaurs are thought to have migrated to North America from Asia, because early relatives first appear on the latter continent. Although there has been some debate about whether tyrannosaurs were active predators or more passive scavengers, the distinction is not usually strong in living predatory animals, and frequently larger carnivores will chase smaller ones away from fresh kills. However, some skeletons of plant-eating dinosaurs evidently have healed wounds caused by tyrannosaur bites, so active predation appears to be sustained.

Ornithomimids were medium-size to large theropods. Almost all of them were toothless, and apparently their jaws were covered by a horny beak; they also had very long legs and arms. A well-known example is *Struthiomimus.* Most were ostrich-sized and were adapted for fast running, with particularly long foot bones, or metatarsals. The largest was *Deinocheirus* from Asia, known only from one specimen consisting of complete arms and hands almost 3 metres (10 feet) long—nearly four times longer than those of *Struthiomimus.* These animals' speed, toothlessness, and long hands with relatively symmetrical fingers leave their lifestyle and feeding habits unclear, but they may have been fairly omnivorous like ostriches, although they are not directly related.

Oviraptorids, therizinosaurids, and caenagnathids appear to form a group slightly more related to birds than to the coelurosaurs. Oviraptorids, known from the Late Cretaceous of Mongolia, had very strange skulls, often with high crests and a reduced dentition in an oddly curved jaw. The name oviraptor means "egg stealer," and it was given because remains of this carnivorous dinosaur were found along with fossil eggs presumed to belong to a small ceratopsian, *Protoceratops,* which lay nearby. Recent discoveries in Mongolia of oviraptorids sitting in birdlike positions on nests of eggs formerly thought to belong to *Protoceratops* reveal that the parentage was misplaced and that oviraptorids, like their bird relatives, apparently tended their young. Therizinosaurids, or segnosaurs, and caenagnathids are known from only a few specimens.

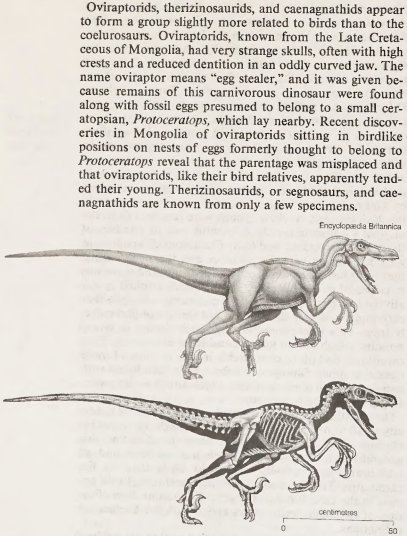


Figure 8: (Top) Rendering of *Velociraptor* and (bottom) reconstruction of its skeleton.

The maniraptorans comprise birds, dromaeosaurs, and troodontids. Dromaeosaurs were medium-size predators with long, grasping arms and hands, moderately long legs, and a specialized stiffened tail that could be used for active balance control. Their feet bore large talons on one toe that were evidently used for raking and slicing prey. A famous discovery known as the "fighting dinosaurs of Mongolia" features a small dromaeosaur, *Velociraptor,* locked in petrified combat with a small protoceratopsian. The hands of the dromaeosaur are grasping the beaked dinosaur's frill, and the foot talons are apparently lodged in its throat. The best-known examples are *Deinonychus* of North America and *Velociraptor* of Asia (Figure 8).

ORNITHISCHIA

The Ornithischia were all plant eaters, as far as is known. In addition to a common pelvic structure, they share a number of other unique features, including a bone that joined the two lower jaws and distinctive leaf-shaped teeth crenulated along the upper edges. They had at least one palpebral, or "eyelid," bone, reduced skull openings near the eyes and in the lower jaw (antorbital and mandibular), five or more sacral vertebrae, and a pubis whose main shaft points backward and down, parallel to the ischium. The earliest and most basal form is the incompletely known *Pisanosaurus,* from the Late Triassic of Argentina. Other primitive forms also existed, but the two main ornithischian lineages are the Cerapoda and Thyreophora.

Cerapoda. Cerapoda is divided into three groups: Ornithopoda, Pachycephalosauria, and Ceratopsia. The latter two are sometimes grouped together as Marginocephalia because they share a few features, including a bony shelf on the back of the skull.

Ornithopoda. Ornithopods include heterodontosaurs, known from southern Africa; the slightly larger hypsilophodontids, about three metres in length; the much larger iguanodontids, about nine metres long, mostly from North America and Europe; and the large duck-billed hadrosaurs of North America and Eurasia.

The posteranial anatomy of the ornithopods reflects the bipedal ancestry of the group, but the giant hadrosaurs and some iguanodontids may have been as comfortable on four legs as on two, especially while feeding on low vegetation. All members had hind legs that were much longer and sturdier than their forelegs. The thighbone (femur) was nearly always shorter than the shinbones (tibia and fibula), especially in all but the largest forms. The tail was long and sometimes quite deep and flat-sided. The vertebral spines of the tail and trunk region were reinforced by a rhomboidal latticework of bony (ossified) tendons running in criss-cross fashion between adjacent spines. They suggest a certain degree of stiffening of the tail and backbone, which were balanced over the massive hips.

Ornithopod vs. theropod feet

Ornithopod feet were modified from the primitive five-toed pattern in a way that resembled similar modifications in theropod feet. The three middle toes served as the functional foot; the inside toe was shortened and often held off the ground, and the outside toe was greatly reduced or absent altogether. The resemblance to theropod feet is so strong that the footprints of the two groups are easily confused, especially if poorly preserved. The toes of all but the most basal ornithopods terminated in broad, almost hooflike bones, especially in the duckbills, as opposed to the sharp claws of theropods, and this is one way to distinguish their footprints. The hand reflected the primitive five-digit design, and, as was generally true in archosaurs, the fourth and fifth digits were shorter than the other three, with the third being longest. In iguanodontids and hadrosaurs, the fingers ended in broad, blunt bones rather than in claws, much like the toes. It is thought that these middle fingers and toes were covered by blunt, hooflike structures. In the duckbills the fingers apparently were encased in a mittenlike structure that could have broadened the hand for better support of the animal's weight on soft ground.

The Ornithopoda differ from one another mainly in the structure of their skulls, their jaws and teeth, their hands and feet, and their pelvises. Ornithopods constitute an excellent case study in evolution because, as the various lin-

eages arise and die out from the latest Triassic to the latest Cretaceous, trends in size, complications and elaborations of teeth and chewing mechanisms, adaptations for quadrupedal posture in some forms, and other changes emerge clearly from their phylogenetic patterns.

In the fabrosaurids the teeth were simple leaf-shaped, laterally compressed elements arranged in a single front-to-back row in each jaw. They were not set in from the outer cheek surface as in most ornithopods. Upper and lower teeth alternated in position when the jaw was closed; they did not occlude directly.

In heterodontosaurs the cheek teeth were crowded together into long rows and set inward slightly from the outer cheek surface. The inset, which persisted through all later ornithopods, has been interpreted to suggest the presence of cheeks that may have held plant food in the mouth for further processing by the cheek teeth. They occluded directly to form distinct chisel-like cutting edges with a self-sharpening mechanism maintained by hard enamel on the outer side of the upper teeth and the inner side of the lower. There were prominent upper and lower tusklike teeth at the front of the mouth (the upper set in the premaxillary bones, the lower on the dentary bones). At least two pairs of incisors seem to have been retained. Certain features of the skull suggest much larger jaw muscles in heterodontosaurs than in the fabrosaurids.

The hypsilophodonts had cheek teeth arranged in tightly packed rows set well inward from the outer cheek surfaces. The teeth occluded directly, and the opposing rows formed a long shearing edge similar to that of the heterodontosaurs. There was, however, no "tusk" either above or below. The premaxillaries had small simple incisor-like teeth above the beak-covered, toothless predentary. Strong projections of bone extended up from the lower jaw toward the moderate-size upper temporal fenestrae.

The skulls of iguanodonts accommodated still larger jaw muscles, but the cheek teeth were less regular and compacted than in the primitive ornithopods and consequently did not occlude as uniformly. Both the premaxillaries and the predentary were toothless but probably were sheathed in horny beaks.

Hadrosaur teeth

Specialization of the teeth and jaws reached a pinnacle in the hadrosaurs, or duck-billed ornithopods. In this group a very prominent, robust projection jutted from the back of the stout lower jaw. Large chambers housing muscles were present above this process and beneath certain openings in the skull (the lateral and upper temporal fenestrae). These chambers are clear evidence of powerful jaw muscles. The dentition consisted of numerous tightly compacted teeth crowded into large grinding batteries. The battery in each jaw was composed of as many as 200 functional and replacement teeth with distinct, well-defined wear, or grinding, surfaces that resulted from very exact occlusion. As teeth were lost from the front of the jaws in iguanodontids and hadrosaurs, the snouts expanded into a bulbous shape, especially in the "duck-billed" hadrosaur, and may have been covered by a horny beak that improved feeding. These bills apparently had edges sharp enough to shred and strip leaves or needles from low shrubs and branches. Pine needles have been identified in duck-billed dinosaur remains and presumably represent stomach contents.

Other interesting specializations may have assisted iguanodontids and hadrosaurs in feeding. The hands were unusually modified in the two groups, though in different ways. In iguanodontids the wrist bones were coalesced into a single blocky structure that was less mobile than in more primitive wrist configurations. The joints of the thumb were similarly coalesced into a single conelike spike that had limited mobility on the wrist. The middle three digits flexed in the normal way and bore broad flat, spatulate claws. The fifth digit actually had two additional joints and became somewhat opposable to the rest of the hand. It is thought that the hands may have been adapted to grasp and strip vegetation, and the spikelike thumb has been suggested to have been an effective weapon against predators. These features were more or less continued in hadrosaurs, except in this group the blocky wrist was reduced and the thumb was lost completely.

Some varieties of hadrosaurs are also noted for the peculiar crests and projections on the top of the head (Figure 9). These structures were expansions of the skull composed almost entirely of the nasal bones. In genera such as *Corythosaurus, Lambeosaurus, Parasaurolophus* (and a few others), the crests were hollow, containing a series of middle and outer chambers that formed a convoluted passage from the nostrils to the trachea. Except for passing air along to the lungs, the function of these crests is not widely agreed upon. Sound production (honking), an improved sense of smell, and a visually conspicuous ornament for species recognition are some suggestions. Because these animals are no longer considered to have been amphibious, ideas such as snorkeling and extra air storage space have generally been discarded. Besides, the crests had no opening at their ends and consequently would not have been able to work as snorkels; even the largest crests held only an estimated 2 percent of the volume of the lungs, hardly enough to justify the construction of such an elaborate structure.

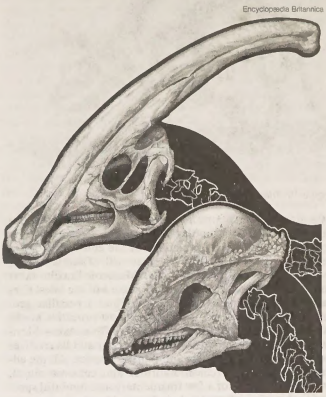


Figure 9: Skulls of a hadrosaur (*Parasaurolophus*) and a pachycephalosaur (*Stegoceras*).

Pachycephalosauria. In important respects the pachycephalosaurs conformed to the basic ornithopod body plan, and there is some evidence that pachycephalosaurs actually evolved from (and are therefore members of) ornithopods, perhaps similar to hypsilophodontids. All of them appear to have been bipedal. They bore the typical ornithopod ossified tendons along the back, and they had simple leaf-shaped teeth, although the teeth were enameled on both sides. The ornithischian type of pelvis was present, but a portion of the ischium was not.

The pachycephalosaurs are known as domeheads because of their most distinctive feature—a marked thickening of the frontoparietal (forehead) bones of the skull (Figure 9). The thickness of bone was much greater than might be expected in animals of their size. The suggestion has been made that this forehead swelling served as protection against the impact of the type of head-butting activities seen today in animals such as bighorn sheep, but microscopic studies of the bone structure of these thick domes suggest that they are poorly designed to divert stresses away from the braincase. Also, the great variety of pachycephalosaur domes—from thin, flat skull tops to pointed ridges with large spikes and knobs facing down and back—suggests no single function in defense or combat.

Stegoceras and *Pachycephalosaurus* of the North American Cretaceous were, respectively, the smallest and largest members of the group, the former attaining a length of about 2.5 metres (8 feet) and the latter twice that. Pachycephalosaurs are known almost entirely from the Late Cre-

taceous (although *Yaverlandia* is from the Early Cretaceous) and have been found in North America and Asia. They are generally rare and still are relatively poorly known among dinosaur groups.

Ceratopsia. The first ceratopsian ("horn-faced") dinosaur remains were found in the 1870s by the American paleontologist Edward D. Cope, who named the animal *Agathaumus,* but the material was so fragmentary that its unusual design was not at once recognized. The first inkling that there had been horned dinosaurs did not emerge until the late 1880s with the discovery of a large horn core, first mistaken for that of a bison. Shortly afterward, dozens of large skulls with horns were found—the first of many specimens of *Triceratops* (Figure 10).

Courtesy, Library Services Department, American Museum of Natural History, New York City; photograph, E.M. Fulda (Neg. No. 310434)

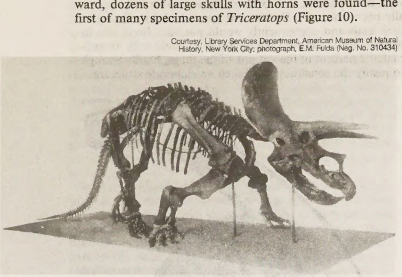


Figure 10: *Triceratops* skeleton.

Ceratopsians first appeared in the modest form of psittacosaurids, or parrot-reptiles, in the Early Cretaceous and survived to the "great extinction" at the end of the Cretaceous Period. *Triceratops,* together with *Tyrannosaurus,* was one of the very last of all known Mesozoic Era dinosaurs in North America, where the fossil record of the latest Cretaceous is best known. Ceratopsians had a peculiar geographic distribution: the earliest and most primitive kinds, such as *Psittacosaurus,* are known only from Asia—Mongolia and China, specifically. *Protoceratops* and its relatives are known from both Asia and North America. All the advanced ceratopsids (chasmosaurines and centrosaurines), with the exception of a few fragmentary and doubtful specimens, have been found only in North America.

Ceratopsians ranged in size from relatively small animals the size of a dog to the nearly 9-metre- (30-foot-) long, four- to five-ton *Triceratops.* Although commonly compared to the modern rhinoceros, *Triceratops* grew to a weight and bulk several times that of the largest living rhinoceros, and its behaviour probably was correspondingly different. The most distinctive feature of nearly all members of the group was the horns on the head, hence the name *ceratops.* Correlated with the various arrays of head horns in the different taxa was the unusually large size of ceratopsian heads. Great bony growths extended from the back of the skull, reaching well over the neck and shoulders. This neck shield, or frill, resulted in the longest head that ever adorned any land animal; the length of the *Torosaurus* skull was almost 3 metres (10 feet), longer than a whole adult *Protoceratops.*

Purpose of the neck frill

Several hypotheses have been proposed to explain this frill structure: a protective shield to cover the neck region, an attachment site of greatly enlarged jaw muscles, an attachment site of powerful neck muscles for wielding the head horns, or a sort of ornament to present a huge, frightening head-on profile to potential attackers. The most unusual thought is that the structure was none of these but rather acted as a giant heat-control apparatus, with its entire upper surface covered in a vast network of blood vessels pulsing with overheated blood or absorbing solar heat.

Most of these hypotheses are difficult to test. One important fact to keep in mind was that the frill was little more than a frame of bone, sometimes ornamented with knobs and spikes around large openings behind and above the skull. An exception to this pattern was *Triceratops,* which had a solid and relatively short frill, but *Triceratops* is so well known that its frill is often mistakenly considered typical of ceratopsians. The open frill of other ceratopsians would have provided only poor protection for the neck region and only a modest area of attachment for jaw or neck muscles. If skin and soft tissues spanned the area framed by the bony frill, it would have created a formidable presence when the head was lowered in threatening display. Such a large structure would naturally have absorbed and reflected sunlight that warmed the tissue and its internal blood vessels, but it is questionable whether this was an important or necessary function of the frill, since other dinosaurs do not have similar structures.

The Ceratopsia are divided into groups that mirror their evolutionary trends through time: the primitive psittacosaurids, such as *Psittacosaurus,* the protoceratopsids, including *Protoceratops* of Asia and *Leptoceratops* of North America; and the ceratopsids, encompassing all the advanced and better-known kinds such as the chasmosaurines *Triceratops* and *Torosaurus* as well as the centrosaurines such as *Centrosaurus* (or *Monoclonius*)—all from North America.

Like the pachycephalosaurs, the most basal ceratopsians, such as *Psittacosaurus,* look much like typical ornithopods, largely because of their relatively long hind limbs and short front limbs (probably resulting in bipedal stance and locomotion) and the persistence of upper front teeth and a fairly unspecialized pelvis. *Psittacosaurus,* however, possessed a beak, the beginnings of a characteristic neck frill at the back of the skull, and teeth that prefigured those of the more advanced ceratopsians. It is also recognized diagnostically as a ceratopsian by the presence of a unique bone called the rostral, a toothless upper beak bone that opposed the lower predentary found in all ornithischians.

The best-known of the protoceratopsids is the genus *Protoceratops.* Dozens of skeletal specimens, ranging from near hatchlings to full-size adults, have been found and studied. This rare treasure, the first to include very young individuals unmistakably associated with mature individuals, was the result of the series of American Museum of Natural History expeditions in the 1920s to the Gobi Desert of Mongolia. Their collection provided the first valid growth series of any dinosaur. Their discovery of several nests of eggs loosely associated with *Protoceratops* skeletons was the first finding of eggs that were unquestionably dinosaurian; originally attributed to *Protoceratops,* the eggs only recently were correctly attributed to the theropod *Oviraptor* (as noted in the section Tetanurae).

The skeletal anatomy of the protoceratopsids foreshadowed that of the more advanced ceratopsids. The ceratopsian skull was disproportionately large for the rest of the animal, constituting about one-fifth of the total body length in *Protoceratops* and at least one-third in *Torosaurus.* The head frill of *Protoceratops* was a modest backward extension of two cranial arches, but it became the enormous fan-shaped ornament of later forms, including *Triceratops.* The advanced ceratopsids are sometimes divided into centrosaurines, which had a prominent nose horn but small or absent eye horns, and chasmosaurines, which had larger eye horns but reduced nose horns.

Ceratopsian jaws

Ceratopsian jaws were highly specialized. The lower jaw was massive and solid to support a large battery of teeth similar to those of the duckbills. The lower jawbones were joined at the front and capped by a stout beak formed of the toothless predentary bone. This structure itself must have been covered by a sharp, horny, turtlelike beak. Continuous dental surfaces extended over the rear two-thirds of the jaw. The tooth batteries, however, differed from those of the hadrosaurs in forming long, vertical slicing surfaces as upper and lower batteries met, operating much like self-sharpening shears.

As in the hadrosaurs, each dental battery consisted of about two dozen or more tooth positions compressed together into a single large block. At each tooth position there was one functional, or occluding, tooth (the duckbills had two or three) along with several more unerupted replacement teeth beneath. (All toothed vertebrates, living and extinct, except mammals, have a lifelong supply of re-

placement teeth.) The suggestion is that they fed on something exceedingly tough and fibrous, such as the fronds of palms or cycads, both of which were plentiful during late Mesozoic times.

With the exception of the bipedal *Psittacosaurus,* and perhaps the facultatively bipedal protoceratopsids, all ceratopsians were obligate quadrupeds with a heavy, ponderous build. The leg bones were stout and the legs themselves muscular; the feet were semiplantigrade for graviportal stance and progression; and all the toes ended in "hooves" rather than claws. As in most other four-legged animals, the rear legs were significantly longer than the front legs (which again suggests their bipedal ancestry). The hind legs were positioned directly beneath the hip sockets and held almost straight and vertical. The front legs, on the other hand, projected out to each side from the shoulder sockets in a "push-up" position. Consequently, the head was carried low and close to the ground. This mixed posture was perhaps related to the large horned head and its role in combat, the bent forelegs providing a wide stance and stable base for directing the horns at an opponent and resisting attack.

The first four neck vertebrae of ceratopsians were fused (co-ossified), presumably to support the massive skull. The first joint of the neck was unusual in that the bone at the base of the skull formed a nearly perfect sphere that fit into a cuplike socket of the fused neck vertebrae. Such an arrangement would seem to have provided solid connections along with maximum freedom of the head to pivot in any direction without having to turn the body. Presumably ceratopsians used their horns in an aggressive manner, but whether they used them as defense against possible predators, in rutting combat with other male ceratopsians, or in both is not so clear. Evidence of puncture wounds in some specimens suggests rutting encounters, but the fact that both sexes apparently had horns seems to indicate defense or species recognition as their primary uses.

Thyreophora. The Thyreophora consist mainly of the well-known Stegosauria, the plated dinosaurs, and Ankylosauria, the armoured dinosaurs, as well as their more basal relatives, including *Scutellosaurus* and *Scelidosaurus.* Both possessed small bony plates, or scutes, along the body.

In the Middle and Late Jurassic, the first stegosaurs and ankylosaurs appeared. Like the previously described forms, they are distinguished by bony scutes. Scutes are maintained and elaborated all over the body in ankylosaurs but are reduced to a series of plates and spikes along the backbone in stegosaurs, though their basic structure remains the same in both groups. Thyreophorans also have low, flat skulls, simple S-shaped tooth rows with small leaf-shaped tooth crowns, and spout-shaped snouts.

Stegosauria. With their unique bony back plates, the stegosaurs are very distinctive. Relatively few specimens have been found, but they were widespread, with remains being found in North America, Africa, Europe, and Asia. Stegosaurian remains have appeared in Early Jurassic to Early Cretaceous strata. The most familiar genus is *Stegosaurus,* found in the Morrison Formation (Late Jurassic) of western North America. *Stegosaurus* was 3.7 metres (12 feet) in height and 9 metres in length, probably weighed two tons, and had a broad, deep body (Figure 11). Not all varieties of the Stegosauria were this large; for example, *Kentrosaurus,* from eastern Africa, was less than 2 metres high and 3.5 metres long.

All stegosaurs were graviportal and undoubtedly quadrupedal, although the massive legs were of greatly disparate lengths—the hind legs being more than twice the length of the forelegs. Whatever walking and running skills were possessed by the stegosaurs, their limb proportions must have made these movements extremely slow. The humerus of the upper arm was longer than the bones of the forearm, the femur much longer than the shinbones, and certain bones of the feet very short, which means that the stride must have been short. In addition, the feet were graviportal in design and showed no adaptations for running.

The stegosaurian skull was notably small, long, low, and narrow, with little space for sizable jaw muscles. The weakly developed dentition consisted of small, laterally compressed, leaf-shaped teeth arranged in short, straight rows. This combination of features seems odd in comparison with the large, bulky body. The weak dentition suggests that the food eaten must have required little preparation by the teeth and yet provided adequate nourishment. Perhaps the digestive tract contained fermenting bacteria capable of breaking down the cellulose-rich Jurassic plant tissues. Digestion may also have been assisted by a crop or gizzard full of pulverizing stomach stones (gastroliths), though none has yet been discovered in stegosaurian specimens. A collection of disklike bones is found in the throat region of *Stegosaurus,* but these are likely to have been embedded in the skin, not used in the gut. Even so, it is still difficult to understand how these animals, with such small and poorly equipped mouths, could have fed themselves adequately to sustain their great bulk. The same problem has been encountered in speculations about the feeding habits of sauropods.

The most distinctive stegosaurian feature was the double row of large diamond-shaped bony plates on the back. A controversy as to their purpose and how they were arranged has raged ever since the first *Stegosaurus* specimen was collected in 1877 by Marsh's workers at the Morrison Formation in Colorado, U.S. The evidence and a general consensus argue in favour of the traditional idea that the plates projected upward and were set in two staggered (alternating) rows on either side of the backbone. In other stegosaurs, such as *Kentrosaurus,* the plates are more symmetrical and may have been arranged side by side. The suggestion that the plates did not project above the back at all, but lay flat to form flank armour, has been rejected on the basis of studies of the microstructure of the bone of the plates, in which attachment fibres are embedded in a manner consistent with an upright position. In *Stegosaurus* itself, the end of the tail bore at least two pairs of long bony spikes, which suggests some sort of defensive role for the tail but not necessarily for the back plates. However, other stegosaurs, such as *Kentrosaurus,* had relatively small plates along the front half of the spine and spikes along the back half of the spine and the tail.

Purpose of the dorsal plates

The discovery in 1976 that the bony plates of *Stegosaurus* were highly vascularized led to the suggestion that these "fins" functioned as cooling vanes to dissipate excess body heat in much the same way that the ears of elephants do. The staggered arrangement in parallel rows might have maximized the area of cooling surface by minimizing any downwind "breeze shadow" that would have resulted from a paired configuration. Asymmetry is a bizarre anatomic condition, and, right or wrong, this certainly is an imaginative explanation of its presence in this animal. No other stegosaur, however, had such a peculiar feature. Rather, all other taxa had a variety of paired body spikes that seem best explained as passive defense or display adaptations rather than cooling mechanisms.

Ankylosauria. The ankylosaurs are known from the Late Jurassic and Cretaceous periods. They are called "armoured dinosaurs" for their extensive mosaic of small and large interlocking bony plates that completely encased the back and flanks. Most ankylosaurs, such as *Euoplocephalus, Nodosaurus,* and *Palaeoscincus,* were relatively

Courtesy, Library Services Department, American Museum of Natural History, New York City; photograph, O. Bauer and R. Sheridan (Neg. No. 2A 13019)

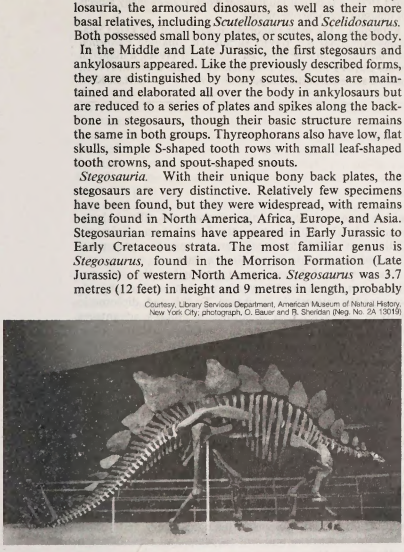


Figure 11: *Stegosaurus* skeleton.

low and broad in body form and walked close to the ground on short, stocky legs in a quadrupedal stance. As in stegosaurs, the hind legs were longer than the front legs, but they were not as disproportionate as those of *Stegosaurus*. Like the stegosaurs, however, their limbs were stout and columnar, the thighbone and upper arm were longer than the shin and forearm, and the metapodials were stubby. These features point to a slow, graviportal mode of locomotion. The feet were semiplantigrade and possibly supported from beneath by pads of cartilage. The bones at the ends of the digits (terminal phalanges) were broad and hooflike rather than clawlike.

The ankylosaur skull was low, broad, and boxlike, with dermal scutes (osteoderms) that were often fused to the underlying skull bones. In *Euoplocephalus* even the eyelid seems to have developed a protective bony covering. The jaws were weak, with a very small predentary and no significant projections of bone for jaw muscle attachment. The small jaw muscle chamber was largely covered by dermal bones rather than having openings. The teeth were small, loosely spaced, leaf-shaped structures reminiscent of the earliest primitive ornithischian teeth. All taxa had very few teeth in either jaw, in marked contrast to the highly specialized, numerous teeth of other ornithischians. These features of the jaws and teeth lead to the impression that the animals must have fed on some sort of soft, pulpy plant food.

Apparently neither very diverse nor abundant, the ankylosaurs are known only from North America, Europe, and Asia. They are divided into the more basal Nodosauridae and the more advanced Ankylosauridae, which may have evolved from nodosaurs. The most conspicuous difference between the two groups is the presence of a massive bony club at the end of the tail in the advanced ankylosaurs; no such tail structure is present in the nodosaurs. The patterns of the armour also generally differ between the two groups, and ankylosaurids tend to have even broader, more bone-encrusted skulls than did the nodosaurs.

BIBLIOGRAPHY

General works. PHILIP J. CURRIE and KEVIN PADIAN (eds.), *Encyclopedia of Dinosaurs* (1997), comprises articles on topics related to dinosaur taxonomy, biology, and evolution as well as important paleontological sites and exhibits worldwide. JAMES O. FARLOW and M.K. BRETT-SURMAN (eds.), *The Complete Dinosaur* (1997), emphasizes aspects of various groups of dinosaurs and their biology. DAVID NORMAN and JOHN SIBBICK, *The Illustrated Encyclopedia of Dinosaurs: An Original and Compelling Insight into Life in the Dinosaur Kingdom* (1985, reissued 1998), provides a well-written and lavishly illustrated treatment that is excellent for the specialist and nonspecialist alike.

Advanced textbooks on vertebrate evolution and paleontology include MICHAEL J. BENTON, *Vertebrate Palaeontology*, 2nd ed. (1997, reissued 2000); and ROBERT L. CARROLL, *Vertebrate Paleontology and Evolution* (1988). DAVID B. WEISHAMPEL, PETER DODSON, and HALSZKA OSMÓLSKA (eds.), *The Dinosauria* (1990), primarily contains extensive reviews of the major taxonomic groups, defining them via anatomic descriptions and drawings while also supplying fossil-site information.

The search for dinosaurs. LOUIE PSIHOYOS and JOHN KNOEBBER, *Hunting Dinosaurs* (1994), assembles an impressive photographic record of the discoveries and the people responsible for them. PHILIPPE TAQUET, *Dinosaur Impressions: Postcards from a Paleontologist* (1998; originally published in French, 1994), vividly traces one paleontologist's travels throughout the world over a period of 30 years. EDWIN H. COLBERT, *Dinosaurs: Their Discovery and Their World* (1961), a landmark treatment of the subject by a world authority of the period, includes extensive photographic and line-drawing coverage, and *Men and Dinosaurs: The Search in Field and Laboratory* (1968, reissued 1971), provides a thorough illustrated history of the discovery, collection, and study of dinosaurs. JOHN R. HORNER and JAMES GORMAN, *Digging Dinosaurs* (1988, reprinted 1995), is a fascinating account of the search for and collecting of dinosaur eggs and nests as told by the discoverers. JOHN H. OSTROM and JOHN S. MCINTOSH, *Marsh's Dinosaurs: The Collections from Como Bluff* (1966, reissued 1999), is illustrated for technical professionals and contains a historical study of one of the most famous dinosaur localities. For specific information about the origin of birds from theropod dinosaurs, LOWELL DINGUS and TIMOTHY ROWE, *The Mistaken Extinction: Dinosaur Evolution and the Origin of Birds* (1998), is an excellent reference source.

Natural history. JOHN R. HORNER and EDWIN DOBB, *Dinosaur Lives: Unearthing an Evolutionary Saga* (1997), explores developments in the understanding of dinosaurian paleobiology. PETER DODSON, *The Horned Dinosaurs: A Natural History* (1996), presents a case history of the ceratopsians, an important group of dinosaurs. KENNETH CARPENTER, KARL F. HIRSCH, and JOHN R. HORNER (eds.), *Dinosaur Eggs and Babies* (1994, reissued 1996), a technical multiauthor work, reviews many aspects of dinosaur reproductive biology.

Extinction. The most authoritative account of the Late Cretaceous extinctions is J. DAVID ARCHIBALD, *Dinosaur Extinction and the End of an Era: What the Fossils Say* (1996), a masterful book. WALTER ALVAREZ, *T. rex and the Crater of Doom* (1998), offers a somewhat different point of view. (J.H.O./K.P.)

Diplomacy

Diplomacy is the established method of influencing the decisions and behaviour of foreign governments and peoples through dialogue, negotiation, and other measures short of war or violence. Modern diplomatic practices are a product of the post-Renaissance European state system. Historically, diplomacy meant the conduct of official (usually bilateral) relations between sovereign states. By the 20th century, however, the diplomatic practices pioneered in Europe had been adopted throughout the world, and diplomacy had expanded to cover summit meetings and other international conferences, parliamentary diplomacy, the international activities of supranational and subnational entities, unofficial diplomacy by nongovernmental elements, and the work of international civil servants.

The term diplomacy is derived via French from the ancient Greek *diplōma*, composed of *diplo*, meaning "folded in two," and the suffix *-ma*, meaning "an object." The folded document conferred a privilege—often a permit to travel—and the term came to denote documents through which princes granted such favours. Later it applied to all solemn documents issued by chancelleries, especially those containing agreements between sovereigns. Diplomacy later became identified with international relations, and the direct tie to documents lapsed (except in diplomatics, which is the science of authenticating old official documents). In the 18th century, the French term *diplomate* ("diplomat" or "diplomatist") came to refer to a person authorized to negotiate on behalf of a state.

The purpose of diplomacy is to strengthen the state, nation, or organization it serves in relation to others by advancing the interests in its charge. To this end, diplomatic activity endeavours to maximize a group's advantages without the risk and expense of using force and preferably without causing resentment. It habitually, but not invariably, strives to preserve peace; diplomacy is strongly inclined toward negotiation to achieve agreements and resolve issues between states. Diplomacy normally seeks to develop goodwill toward the state it represents, nurturing relations with foreign states and peoples that will ensure their cooperation or—failing that—their neutrality. Even in times of peace, however, diplomacy may involve coercive threats of economic or other punitive measures or demonstrations of the capability to impose unilateral solutions to disputes by the application of military power. When diplomacy fails, war may ensue. Nevertheless, diplomacy is useful even during war. It conducts the passages from protest to menace, dialogue to negotiation, ultimatum to reprisal, and war to peace and reconciliation with other states.

This article discusses the history of diplomacy and the ways in which modern diplomacy is conducted. For a discussion of the legal rules governing diplomatic negotiation and the preparation of treaties and other agreements, see INTERNATIONAL LAW. The United Nations, one venue for diplomacy, is considered in detail in UNITED NATIONS.

Foreign policy and diplomatic history in the 20th century is discussed in INTERNATIONAL RELATIONS, 20TH-CENTURY.

For coverage of related topics in the *Macropædia* and *Micropædia*, see the *Propædia*, section 544, and the *Index*.

The article is divided into the following sections:

History of diplomacy 331
The ancient world
The Middle Ages
The Renaissance to 1815
The Concert of Europe to World War I
The 1920s to the 1980s
The end of the Cold War
Modern diplomatic practice 337
Diplomatic agents
Rights and privileges
Credentials
Diplomatic tasks
Diplomatic agreements
Conference diplomacy
Personnel
Bibliography 340

HISTORY OF DIPLOMACY

The ancient world. Some elements of diplomacy predate recorded history. Early societies had some attributes of states, and the first international law arose from intertribal relations. Tribes negotiated about marriages and regulations on trade and hunting. Messengers and envoys were accredited, sacred, and inviolable; they usually carried some emblem, such as a message stick, and were received with elaborate ceremonies. Women often were used as envoys because of their perceived mysterious sanctity and their use of "sexual wiles"; it is believed that women regularly were entrusted with the vitally important task of negotiating peace in primitive cultures.

Egyptian diplomacy

Information regarding the diplomacy of early peoples is based on sparse evidence. There are traces of Egyptian diplomacy dating to the 14th century BC, but none has been found in western Africa before the 9th century AD. The inscriptions on the walls of abandoned Mayan cities indicate that exchanges of envoys were frequent, though almost nothing is known of the substance or style of Mayan and other pre-Columbian Central American diplomacy. In South America the dispatch of envoys by the expanding Incan Empire appears to have been a prelude to conquest rather than an exercise in bargaining between sovereigns.

Our greatest knowledge of early diplomacy concerns eastern West Asian and Mediterranean peoples, the ancient Chinese, and the Indians. Records of treaties between Mesopotamian city-states date from about 2850 BC. Thereafter, Akkadian (Babylonian) became the first diplomatic language, serving as the international tongue of the Middle East until it was replaced by Aramaic. A diplomatic correspondence from the 14th century BC existed between the Egyptian court and a Hittite king on cuneiform tablets in Akkadian—the language of neither. The oldest treaties of which full texts survive, from about 1280 BC, were between Ramses II of Egypt and Hittite leaders. There is significant evidence of Assyrian diplomacy in the 7th century and of the relations of Jewish tribes, found chiefly in the Bible, with each other and other peoples.

China. The first records of Chinese and Indian diplomacy date from the 1st millennium BC. By the 8th century BC, the Chinese had leagues, missions, and an organized system of polite discourse between their many "warring states." The Chinese tradition was very sophisticated, emphasizing the practical virtues of ethical behaviour in relations between states, and is well documented in the Chinese classics. The Chinese tradition of equal diplomatic dealings between contending states within China was ended by the country's unification under the Qin emperor in 221 BC and the consolidation of unity under the Han dynasty in 206 BC. Under the Han and succeeding dynasties, China emerged as the largest, most populous, technologically most advanced, and best-governed society in the world. The arguments of earlier Chinese philosophers, such as Mencius, prevailed; the best way for a state to exercise influence abroad, they had said, was to develop a moral society worthy of emulation by admiring foreigners and to wait confidently for them to come to China to learn.

Once each succeeding Chinese dynasty consolidated its rule at home and established its borders with the non-Chinese world, its foreign relations were typically limited to the defense of China's borders against foreign attacks or incursions, the reception of emissaries from neighbouring states seeking to ingratiate themselves and to trade with the Chinese state, and the control of foreign merchants in specific ports designated for foreign trade. With rare exceptions, Chinese leaders and diplomats waited at home for foreigners to pay their respects rather than venturing abroad themselves. This "tributary system" lasted until European colonialism overwhelmed it and introduced to Asia the European concepts of sovereignty, suzerainty, spheres of influence, and other diplomatic norms, traditions, and practices.

India. Ancient India also possessed an equally sophisticated but very different diplomatic tradition, which was systematized and described in the Artha-śāstra by the scholar-statesman Kautilya at the end of the 4th century BC. The ruthlessly realistic state system codified in the Artha-śāstra insisted that foreign relations be determined by self-interest rather than by ethical considerations. It graded state power with respect to five factors and emphasized espionage, diplomatic maneuver, and contention by 12 categories of states within a complex geopolitical matrix. It also posited four expedients of statecraft (conciliation, seduction, subversion, and coercion) and six forms of state policy (peace, war, nonalignment, alliances, shows of force, and double-dealing). To execute policies derived from these strategic geometries, ancient India fielded three categories of diplomats (plenipotentiaries, envoys entrusted with a single issue or mission, and royal messengers); a type of consular agent, who was charged with managing commercial relations and transactions; and two kinds of spies (those charged with the collection of intelligence and those entrusted with subversion and other forms of covert action).

The Artha-śāstra

The Indian state, which was separated from its neighbours by deserts, seas, and the Himalayas, had little political connection to other regions of the world until Alexander the Great conquered its northern regions in 326 BC. The establishment of the native Mauryan empire ushered in a new era in Indian diplomatic history that attempted to extend both Indian religious doctrines and political influence beyond South Asia. The Mauryan emperor Aśoka was particularly active, receiving several emissaries from the Macedonian-ruled kingdoms and dispatching numerous Brahman-led missions of his own to West, Central, and Southeast Asia. Such contacts continued for centuries until the ascendancy of the Rājput kingdoms (8th to the 13th century AD) again isolated northern India from the rest of the world. Outside the Cōla dynasty and other Dravidian kingdoms of South India, India's distinctive mode of diplomatic reasoning and early traditions were forgotten and replaced by its Muslim and British conquerors.

Greece. The tradition that ultimately inspired the birth of modern diplomacy in post-Renaissance Europe and that led to the present world system of international relations began in ancient Greece. Some of the first traces of interstate relations concern the Olympic Games of 776 BC. In the 6th century BC, the amphictyonic leagues maintained interstate assemblies with extraterritorial rights and permanent secretariats. Sparta was actively forming alliances in the mid-6th century BC, and by 500 BC it had created

Amphictyonic leagues

the Peloponnesian League. In the 5th century BC, Athens led the Delian League during the Greco-Persian Wars.

Greek diplomacy took many forms. Heralds, references to whom can be found in prehistory, were the first diplomats and were protected by the gods with an immunity that other envoys lacked. Because heralds were inviolable, they were the favoured channels of contact in wartime, preceding envoys to arrange for safe passage. Whereas heralds traveled alone, envoys journeyed in small groups, to ensure each other's loyalty. Because they were expected to sway foreign assemblies, envoys were chosen for their oratorical skills. Unlike modern ambassadors, heralds and envoys were short-term visitors in the city-states whose policies they sought to influence.

In marked contrast to diplomatic relations, commercial and other apolitical relations between city-states were conducted on a continuous basis. Greek consular agents, or *proxeni*, were citizens of the city in which they resided, not of the city-state that employed them. Like envoys, they had a secondary task of gathering information, but their primary responsibility was trade.

The Greeks developed archives, a diplomatic vocabulary, principles of international conduct that anticipated international law, and many other elements of modern diplomacy. Their envoys and entourages enjoyed diplomatic immunity for their official correspondence and personal property. Truces, neutrality, commercial conventions, conferences, treaties, and alliances were common. In one 25-year period of the 4th century BC, for example, there were eight Greco-Persian congresses, where even the smallest states had the right to be heard.

Rome. Rome inherited what the Greeks devised and adapted it to the task of imperial administration. As Rome expanded, it often negotiated with representatives of conquered areas, to which it granted partial self-government by way of a treaty. Treaties were made with other states under Greek international law. During Rome's dominance, envoys were received with ceremony and magnificence, and they and their aides were granted immunity.

Roman envoys were sent abroad with written instructions from their government. Sometimes a messenger, or *nuntius*, was sent, usually to towns. For larger responsibilities, a *legatio* ("embassy") of 10 or 12 *legati* ("ambassadors") was organized under a president. The *legati*, who were leading citizens chosen for their skill at oratory, were inviolable. Rome also created sophisticated archives, which were staffed by trained archivists. Paleographic techniques were developed to decipher and authenticate ancient documents.

Roman law became the basis of treaties. Late in the Republican era, the laws applied by the Romans to foreigners and to foreign envoys were merged with the Greek concept of natural law to create a "law of nations." The sanctity of treaties and the law of nations were absorbed by the Roman Catholic church and preserved in the centuries after the Western Roman Empire collapsed in the 5th century AD.

Even as monarchs negotiated directly with nearby rulers or at a distance through envoys from the 5th through the 9th century, the papacy continued to use *legati*. Both forms of diplomacy intensified in the next three centuries. Moreover, the eastern half of the Roman Empire continued for nearly 1,000 years as the Byzantine Empire. Its court at Constantinople, to which the papacy sent envoys from the mid-5th century, had a department of foreign affairs and a bureau to deal with foreign envoys. Aiming to awe and intimidate foreign envoys, Byzantium's rulers marked the arrival of diplomats with spectacular ceremonies calculated to suggest greater power than the empire actually possessed.

Islāmic diplomacy

The Middle Ages. *Islām.* Inspired by their religious faith, followers of Islām in Arabia conquered significant territory beginning in the 7th century, first by taking Byzantium's southern and North African provinces and then by uniting Arabs, Persians, and ultimately Turks and other Central Asian peoples in centuries of occasionally bloody conflict with the Christian Byzantines. Islāmic diplomatic missions, both to other Muslim states and to non-Muslim states, existed from the time of Muḥammad, and early Islāmic rulers and jurists developed an elaborate set of protections and rules to facilitate the exchange of emissaries. As Muslims came to dominate vast territories in Africa, Asia, and Europe, the experience of contention with Byzantium shaped Islāmic diplomatic tradition along Byzantine lines.

Byzantium. Byzantium produced the first professional diplomats, who were issued written instructions and were enjoined to be polite and to encourage trade. From the 12th century, their role as gatherers of information about conditions in their host states became increasingly vital to the survival of the Byzantine state. As its strength waned, timely intelligence from Byzantine diplomats enabled the emperors to play off foreign countries against each other. Byzantium's use of diplomats as licensed spies and its employment of the information they gathered to devise skillful and subtle policies to compensate for a lack of real power inspired many governments. After the Byzantine Empire's collapse in 1453, major elements of its diplomatic tradition lived on in the Ottoman Empire and in Renaissance Italy.

The Roman Catholic church. As Byzantium crumbled, the West revived. Indeed, even in its period of greatest weakness, the Roman Catholic church conducted an active diplomacy, especially at Constantinople and in its 13th-century struggle against the Holy Roman emperors. Popes served as arbiters, and papal legates served as peacemakers. The prestige of the church was such that, at every court, papal emissaries took precedence over secular envoys, a tradition that continues in countries where Roman Catholicism is the official religion. The Roman emphasis on the sanctity of legates became part of canon law, and church lawyers developed increasingly elaborate rules governing the status, privileges, and conduct of papal envoys, rules that were adapted later for secular use.

Papal diplomacy

From the 6th century, both legates and (lesser-ranking) *nuncii* ("messengers") carried letters of credence to assure the rulers to whom they were accredited of the extent of their authority as agents of the pope, a practice later adopted for lay envoys. A *nuncius* was a messenger who represented and acted legally for the pope; *nuncii* could negotiate draft agreements but could not commit the pope without referral. In time, the terms legate and *nuncius* came to be used for the diplomatic representatives of both secular rulers and the pope. By the 12th century, the secular use of *nuncii* as diplomatic agents was commonplace.

When diplomacy was confined to nearby states and meetings of rulers were easily arranged, a visiting messenger like the *nuncius* sufficed. As trade revived, however, negotiations at a distance became increasingly common. Envoys no longer could refer the details of negotiations to their masters on a timely basis. They therefore needed discretionary authority to decide matters on their own. To meet this need, in the 12th century the concept of a procurator with *plena potens* ("full powers") was revived from Roman civil law. This plenipotentiary could negotiate and conclude an agreement, but, unlike a *nuncius*, he could not represent his principal ceremonially. As a result, one emissary was often given both offices.

At the end of the 12th century, the term ambassador appeared, initially in Italy. Derived from the medieval Latin *ambactiare*, meaning "to go on a mission," the term was used to describe various envoys, some of whom were not agents of sovereigns. Common in both Italy and France in the 13th century, it first appeared in English in 1374 in *Troilus and Criseyde* by Geoffrey Chaucer. By the late 15th century, the envoys of secular rulers were commonly called ambassadors, though the papacy continued to send legates and *nuncii*.

The Crusades and the revival of trade increased Europe's contact with the eastern Mediterranean and West Asia. Venice's location provided it with early ties with Constantinople, from which it absorbed major elements of the Byzantine diplomatic system. Venice gave its envoys written instructions and established a systematic archive. Later it developed an extensive diplomacy on the Byzantine model, which emphasized the reporting of conditions in the host country. Initially, returning Venetian envoys presented their *relazione* ("final report") orally, but they

began in the 15th century to present such reports in writing. Other Italian city-states, then France and Spain, copied Venetian diplomatic methods and style.

The Renaissance to 1815. It is unclear which Italian city-state had the first permanent envoy. In the late Middle Ages and the early Renaissance, most embassies were temporary. As early as the late 14th and early 15th centuries, however, Venice, Milan, and Mantua sent resident envoys to each other, to the popes, and to the Holy Roman emperors. Resident embassies became the norm in Italy in the late 15th century, and after 1500 the practice spread northward. A permanent Milanese envoy to the French court of Louis XI arrived in 1463 and was later joined by a Venetian representative. Ambassadors served a variety of roles, including reporting events back home and negotiating with their hosts. In addition, they absorbed the role of commercial consuls, who were not then diplomatic agents.

Growth of resident embassies

Italy's economic revival, geographic location, and small size fostered the creation of a European state system in microcosm. Because the peninsula was fully organized into states, wars were frequent, and the maintenance of a balance of power necessitated constant diplomatic interaction. Whereas meetings of rulers were considered risky, unobtrusive diplomacy by resident envoys was deemed safer and more effective. Thus, the system of permanent agents took root.

Rome became the centre of Italian diplomacy and of intrigue, information gathering, and spying. Popes received ambassadors but did not send them. The papal court had the first organized diplomatic corps: the popes addressed the envoys jointly, seated them as a group for ceremonies, and established rules for their collective governance.

As resident missions became the norm, ceremonial and social occasions came to dominate the relations between diplomats and their hosts. Papal envoys took precedence over those of temporal rulers. Beyond this, there was little agreement, and there was frequent strife. Pope Julius II established a list of precedence in 1504, but it did not solve the problem. Spain did not accept inferiority to France, power fluctuated among the states, papal power declined, and the Protestant Reformation complicated matters—not least regarding the pope's own position. By the 16th century, the title of ambassador was being used only for envoys of crowned heads and the republic of Venice. Latin remained the international language of diplomacy.

The French invasion of Italy in 1494 confronted the Italian states with a power greater than any within their own state system. They were driven to substitute subtle diplomacy and expedient, if short-lived, compromise for the force they lacked. This tendency, plus their enthusiasm for diplomatic nuances and for the 16th-century writings of Niccolò Machiavelli, gave Italian diplomacy a reputation for deviousness. But it was no more so than that of other states, and Machiavelli, himself a diplomat, argued that an envoy needed integrity, reliability, and honesty—views seconded since by virtually every authority.

The 16th-century wars in Italy, the emergence of strong states north of the Alps, and the Protestant Reformation ended the Italian Renaissance but spread the Italian system of diplomacy. Henry VII of England adopted the Italian diplomatic system, and initially he even used Italian envoys. By the 1520s, Thomas Cardinal Wolsey, Henry VIII's chancellor, had created an English diplomatic service. France adopted the Italian system in the 1520s and had a corps of resident envoys by the 1530s.

Because they were highly trusted as personal emissaries of one ruler to another, and because communications were slow, ambassadors enjoyed considerable freedom of action. Their task was complicated by the ongoing religious wars, which generated distrust, narrowed contacts, and jeopardized the reporting that was essential before newspapers were widespread.

The religious wars of the early 17th century were an Austro-French power struggle. During the Thirty Years' War, innovations occurred in the theory and practice of international relations. In 1625 the Dutch jurist Hugo Grotius published *De Jure Belli ac Pacis* (*On the Law of War and Peace*), in which the laws of war were most numerous. In an effort to convert the law of nations into a law among nations and to provide it with a new secular rationale acceptable to both sides in the religious quarrel, Grotius fell back on the classical view of natural law and the rule of reason. He also enunciated the concepts of state sovereignty and the equality of sovereign states.

Grotius' *Law of War and Peace*

The first modern foreign ministry was established in 1626 in France by Armand-Jean du Plessis Cardinal Richelieu. He created the Ministry of External Affairs to centralize policy and to ensure his control of envoys as he pursued the *raison d'état* ("national interest"). Richelieu rejected the view that policy should be based on dynastic or sentimental concerns or a ruler's wishes, holding instead that the state transcended crown and land, prince and people, having interests and needs independent of all these elements. He asserted that the art of government lay in recognizing these interests and acting according to them, regardless of ethical or religious considerations.

Richelieu's practices led him to ally Roman Catholic France with the Protestant powers in the Thirty Years' War against France's great rival, Austria. He largely succeeded, for the Peace of Westphalia of 1648 weakened Austria and enhanced French power. The four years of meetings before its signature were the first great international congresses of modern history. Princes attended, but diplomats did most of the work in secret meetings. The task of the diplomats was complicated by the need for two simultaneous congresses, because the problem of precedence was otherwise insoluble.

The Treaty of Westphalia did not solve precedence disputes. The war between France and Spain, which continued from 1648 to 1659, was partly about this issue. Shortly thereafter, in 1661, there was a diplomatic dispute in London concerning whether the French ambassador's carriage would precede that of his Spanish rival. War was narrowly averted, but questions of precedence continued to bedevil European diplomacy.

Diplomacy of Louis XIV

During Louis XIV's reign, aristocratic envoys became common, not least because of the expense involved. As a result of Louis XIV's preeminence, French superseded Latin as the language of diplomacy and continued as the lingua franca of diplomacy until the 20th century.

Louis XIV personally directed French foreign policy and read the dispatches of his ambassadors himself. The foreign minister belonged to the Council of State and directed a small ministry and a sizable diplomatic corps under the king's supervision. Envoys were assigned for three or four years and given letters of credence, instructions, and ciphers for secret correspondence. Louis XIV's frequent wars resulted at their conclusion in peace congresses, which were attended by diplomats.

Some states regularized the position of consuls as state officials, though they were not considered diplomats. The French system was imitated by other major states in the 18th century. The ambassadors they sent forth were true plenipotentiaries, able to conclude treaties on their own authority. The title of ambassador was used only for the envoys of kings (and for those from Venice). The diplomacy of the time recognized the existence of great powers by according special rank and responsibility to the representatives of these countries. New among these was Russia, whose diplomatic tradition married elements derived directly from Byzantium to the now essentially mature diplomatic system that had arisen in western Europe.

At the century's end, a power of second rank appeared outside Europe: the United States. The founders of American diplomacy (*e.g.*, Benjamin Franklin and Thomas Jefferson) accepted the norms of European diplomacy but declined to wear court dress or to adopt usages they considered unrepublican. To this day, American ambassadors, unlike those of other countries, are addressed not as "Your Excellency" but simply as "Mr. Ambassador."

The Concert of Europe to World War I. *The Concert and the balance of power.* In the 18th century, European diplomacy strove to maintain a balance between Britain, France, Austria, Russia, and Prussia. At the century's end, however, the French Revolution and the attempts of Napoleon I to conquer Europe first unbalanced and then overthrew the continent's state system. After Napoleon's defeat, the Congress of Vienna was convened in 1814–15

The Congress of Vienna

to set new boundaries, re-create the balance of power, and guard against future French hegemony. The Final Act of Vienna of 1815, as amended at the Congress of Aix-la-Chapelle (Aachen) in 1818, established four classes of heads of diplomatic missions—precedence within each class being determined by the date of presentation of credentials—and a system for signing treaties in French alphabetical order by country name. Thus ended the battles over precedence. Unwritten rules also were established. At Vienna, for example, a distinction was made between great powers and "powers with limited interests." Only great powers exchanged ambassadors. Until 1893 the United States had no ambassadors; like other lesser states, its envoys were only ministers.

More unwritten rules were soon developed. Napoleon's return and second defeat required a new peace treaty with France at Paris in November 1815. On that occasion, the four great victors (Britain, Austria, Russia, and Prussia) formally signed the Quadruple Alliance, which called for periodic meetings of the signatories to consult on common interests, to ensure the "repose and prosperity of the Nations," and to maintain the peace of Europe. This clause, which created a Concert of Europe, entailed cooperation and restraint, as well as a tacit code: the great powers would make all important decisions; internal changes in any member of the Concert had to be sanctioned by the great powers; the great powers were not to challenge each other; and the Concert would decide all disputes. The Concert thus constituted a rudimentary system of international governance by a consortium of great powers.

Initially, meetings of the Concert were attended by rulers, chancellors, and foreign ministers. The first meeting (1818) resulted in the admittance of France to the Concert and the secret renewal of the Quadruple Alliance against it. It was the first international congress held in peacetime and the first to attract coverage by the press. Thus was born the public relations aspect of diplomacy and the press communiqué.

Thereafter, congresses met in response to crises. Owing to disputes between the powers, after 1822 the meetings ceased, though the Concert itself continued unobtrusively. Beginning in 1816 an ambassadorial conference was established in Paris to address issues arising from the 1815 treaty with France. Other conferences of ambassadors followed to address specific international problems and to sanction change when it seemed advisable or unavoidable. The Concert was stretched and then disregarded altogether between 1854 and 1870, during the Crimean War and the unifications of Italy and Germany. The century during which it existed (1815–1914) was generally peaceful, marred only by short, limited wars.

Conference diplomacy and the effects of democratization. After three decades, Europe reverted to conference diplomacy at the foreign-ministerial level. The Congress of Paris of 1856 not only ended the Crimean War but also resulted in the codification of a significant amount of international law. As European powers extended their sway throughout the world, colonies and spheres of influence in areas remote from Europe came increasingly to preoccupy their diplomacy. Conferences in Berlin in 1878 and 1884–85 prevented conflagrations over the so-called "Eastern" and "African" questions—euphemisms, respectively, for intervention on behalf of Christian interests in the decaying Ottoman Empire and the carving up of Africa into European-ruled colonies. Furthermore, multilateral diplomacy was institutionalized in a permanent form. The peace conferences at The Hague (1899–1907), which resulted in conventions aimed at codifying the laws of war and encouraging disarmament, were harbingers of the future.

The Hague peace conferences

During the 19th century the world underwent a series of political transformations, and diplomacy changed with it. In Europe, power shifted from royal courts to cabinets. Kings were replaced by ministers at international meetings, and foreign policy became a matter of increasingly democratized politics. With mass literacy and the advent of inexpensive newspapers, foreign policy came to be swayed by public opinion.

The spread of European diplomatic norms. Meanwhile, European culture and its diplomatic norms spread throughout the world. Most Latin American colonies became independent, and they adopted the existing system without question. The British Empire then united all of the subcontinent for the first time under a single sovereignty. In the mid-19th century, an American naval flotilla forced Japan to open its society to the rest of the world. Afterward, Japan embarked on a rapid program of modernization based on the wholesale adoption of Western norms of political and economic behaviour.

In the late 18th and early 19th centuries, European emissaries to China faced demands to prostrate themselves ("kowtow") to the Chinese emperor in order to be formally received by him in Beijing (Peking), a humiliating practice not encountered since the era of Byzantium. As plenipotentiary representatives of foreign sovereigns, they viewed it as completely inconsistent with the Westphalian concept of sovereign equality. The Chinese, for their part, neither understood nor accepted European diplomatic concepts and practices and were vexed and insulted by the unwillingness of Western representatives to respect the long-established ceremonial requirements of the Chinese court. In the ensuing argument over diplomatic protocol, Europeans prevailed by force of arms. In 1860 British and French forces sacked and pillaged the emperor's summer palace and some areas around Beijing. They refused to withdraw until the Chinese court had agreed to receive ambassadors on terms consistent with Western practices and to make other concessions.

Western diplomacy beyond Europe initially was conducted at a leisurely pace, given the vastly greater distances and times required for communication. Fortunately, the dispatch of far-flung legations developed almost simultaneously with advances in transportation and communications, which made frequent contact possible. The railway, the telegraph, the steamship, and undersea cable sped the transmittal of instructions and information. Improvements in technology now made referral to the capital possible and ensured that capitals heard from their envoys abroad, even in the most distant places, on a more frequent and timely basis.

Speedier communication, more involvement in commercial diplomacy as trade became crucial to prosperity, and, especially, the advent of typewriters and mimeograph machines all contributed to a significant increase in the number of diplomatic reports. Yet diplomacy remained a relatively gilded and leisurely profession. It also remained a relatively small one. Before 1914 there were 14 missions in Washington, D.C.; by the beginning of the 21st century, the number of missions had increased by more than 12-fold. Diplomats shared a similar education, ideology, and culture. They saw themselves as an elite and carefully upheld the fiction that they still were personal envoys of one monarch to another.

The 1920s to the 1980s. *The League of Nations and the revival of conference diplomacy.* Despite its risks and inherent complexity, conference diplomacy was revived during World War I and continued afterward, especially during the 1920s. The Paris Peace Conference took place amid much publicity, which was intensified by the newsreels made of the event. U.S. President Woodrow Wilson had enunciated his peace program in January 1918, including "open covenants of peace openly arrived at" as a major goal of diplomacy in the post-World War I period. His phrasemaking, which entangled process and result, caused confusion. Hundreds of journalists went to the conference only to discover that all but the plenary sessions were closed. Wilson had intended that the results of diplomatic negotiations be made public, with treaties published and approved by legislatures. He largely achieved this goal, as the Covenant of the League of Nations required that treaties be registered at the League before they became binding.

The Paris conference and later conferences were conducted in both English and French. As at Vienna, political leaders attended, but kings and princes were strikingly absent. Even more strenuously than at Vienna, nongovernmental organizations, most representing national entities seeking independence, sought a hearing at the court of the great powers. Ultimately, some European peoples gained

independence, which resulted in a dramatic increase in the number of sovereign states.

The League of Nations

The chief innovation of the peace negotiations was the creation of the League of Nations as the first permanent major international organization, with a secretariat of international civil servants. The League introduced parliamentary diplomacy in a two-chamber body, acknowledging the equality of states in its lower house and the supremacy of great powers in its upper one. As neither chamber had much power, however, the sovereignty of members was not infringed. The League of Nations sponsored conferences and supervised specialized agencies, some of which were newly created.

Despite the presence of a Latin American bloc and a few independent or quasi-independent states of Africa and Asia, the League of Nations was primarily a European club. Diplomats became orators again in the halls of Geneva, but the topics were often trivial. Decisions taken in public were rehearsed in secret sessions. On important matters, foreign ministers attending League councils met privately. In 1923 the League revealed its impotence by dodging action in the Corfu crisis, in which Italian troops occupied the Greek island following the murder of an Italian general on Greek soil. In later years the League failed to improve its record on international crises.

The weakness of the League of Nations was aggravated by the absence of the United States, whose Senate refused to ratify the peace treaties by which the League was created. The Senate's inaction raised questions about the country's reliability—the basis of effective diplomacy—and drew attention to the blurring of the line between foreign and domestic policy and, in the view of some, the irresponsibility of democratic electorates. It also rendered good diplomacy—which is based on compromise, mutual advantage, and lasting interests—extremely difficult.

In Europe, where electorates were constantly preoccupied with foreign policy, this problem was most acute. Statesmen, trailed by the popular press, engaged in personal diplomacy at frequent conferences. Foreign offices, diplomats, and quiet negotiation were eclipsed as prime ministers and their staffs executed policy in a blaze of publicity. Governments manipulated this publicity to influence public opinion in favour of their policies. As the masses became concerned with such matters, unprecedented steps were taken to bribe the foreign press, to plant stories, and to use public occasions for propaganda speeches aimed at foreign audiences.

Despite these changes, the "new diplomacy" of the early 20th century was not so new. The negotiating process remained the same. Talks continued to be held in secret, and usually only their results were announced to the public. Meanwhile, diplomats deplored the decline of elite influence and the effects of expanded democracy—*e.g.*, press scrutiny, public attention, and the involvement of politicians—on the diplomatic process.

The tensions of the 1930s revived conference diplomacy, which continued during World War II. Thereafter, summit meetings between heads of government became the norm as technology again quickened the tempo of diplomacy. In the 1930s, statesmen began to telephone each other, a practice that was epitomized in the 1960s by the Soviet-American "hot line." Similarly, the flights of British Prime Minister Neville Chamberlain to Germany in 1938, resulting in the Munich agreement that allowed Germany to annex the Sudetenland in western Czechoslovakia, started a trend in diplomacy. With airplanes at their disposal, leaders met often in the postwar world. As Kojo Debrah, a Ghanaian diplomat, later remarked, "Radio enables people to hear all evil, television enables them to see all evil, and the jet plane enables them to go off and do all evil."

The rise of totalitarian regimes. World War I accelerated many changes in diplomacy. The Russian Revolution of 1917 produced a powerful new regime that rejected the political assumptions and practices of the Western world and used political language—including the terms democracy, propaganda, and subversion—in new ways. The communist government of the Soviet Union abolished diplomatic ranks and published the secret treaties it found in the czarist archives. Without delay, the People's Commissariat of Foreign Affairs (known by its Russian acronym, the Narkomindel) organized a press bureau and a bureau for international revolutionary propaganda. As Russia entered peace negotiations with Germany, it substituted propaganda for the power it lacked, appealing openly to the urban workers of other states to exert pressure on their governments. Under the leadership of Joseph Stalin, the Soviet Union used each concession it won as a basis to press for another, and it viewed diplomacy as war, not as a process of mutual compromise. Nazi Germany was equally indifferent to accommodation and Western opinion once it achieved rearmament; Adolf Hitler signed treaties with the intention of keeping them only as long as the terms suited him, regarded with contempt those who accommodated him, and cowed foreign leaders with tantrums and threats.

Soviet diplomacy

Diplomacy during the Cold War. After World War II, most of the world divided into two tight blocs, one dominated by the United States and one by the Soviet Union. The Cold War took place under the threat of nuclear catastrophe and gave rise to two major alliances—the North Atlantic Treaty Organization, led by the United States, and the Warsaw Pact, led by the Soviet Union—a conventional and nuclear arms race, endless disarmament negotiations, much conference diplomacy, many summits, and periodic crisis management, a form of negotiation aimed at living with a problem, not solving it. As a result, a premium was placed on the diplomatic art of continuing to talk until a crisis ceased to boil. The widening Cold War entailed more espionage, of which ambassadors were officially ignorant; thus, large embassies appeared in small but strategic countries. Propaganda and so-called "cultural diplomacy"—as typified by the international tours of Russian dance companies and the cultural programs of the Alliance Française, the British Council, and various American libraries—expanded as well. Cold War competition also extended to international arms transfers. Gifts or sales of weapons and military training were a means of influencing foreign armed forces and consolidating relationships with key elements of foreign governments. The increasing complexity and expense of modern weapons also made military exports essential for preserving industrial capacity and employment in the arms industries of the major powers. Diplomats thus became arms merchants, selling weapons to their host governments.

The many diplomatic tasks reflected a world that was more fragmented and divided. This dangerous situation led to a search for mechanisms to manage the Cold War in order to prevent a nuclear holocaust. Neither the UN nor the Western policy of containment provided an answer. As the two blocs congealed, a balance of terror in the 1960s was followed by an era of détente in the 1970s and then by a return to deterrence in the 1980s. But the 45 years of Cold War did not produce an organizing principle of any duration. Great-power conflict was conducted by proxy through client states in developing areas.

Diplomats as targets

As old diplomatic premises broke down, diplomacy became a hazardous career. Diplomats were targeted because they represented states and symbolized privileged elites. Security precautions at embassies were increased (some came to resemble fortresses) but were insufficient if host governments turned a blind eye to breaches of extraterritoriality. Attacks on diplomatic missions and diplomats grew, with terrorists succeeding in taking the staffs of some diplomatic missions hostage and in blowing up others.

Some new states also adopted the Soviet tactic of offensive behaviour as a tool of policy, appealing, over the heads of government, to the common people in the opponent's camp. It tried to discredit governments by accusing them of ugly motives, and it sometimes trumpeted maximum demands in calculatedly offensive language as conditions for negotiation. Public diplomacy of this ilk was often noisy, bellicose, and self-righteous.

As diplomacy raised its voice in public, propaganda, abetted by technology, became a key tool. Radio Free Europe and the Voice of America broadcast one message to the communist bloc; proselytizing Christian churches and so-called "national liberation movements" used transistor ra-

dios to spread their messages to other areas. Television also became crucial, as its images provided an immediacy that words alone could not convey. Mass demonstrations were staged for the benefit of television and featured banners in English, which had become the dominant international language. When the United States invaded Panama in 1989, the Soviet Union protested on the American-owned television company Cable News Network, which was watched by most world leaders.

The effects of decolonization. World War II also sounded the death knell for global empires. The immediate postwar period saw the reemergence into full independence of several great civilizations that the age of imperialism had placed under generations of European tutelage. Many of these reborn countries adopted with zeal the central doctrines of European diplomacy, including the concepts of sovereignty, territorial integrity, and noninterference in internal affairs.

After a long struggle for independence, Indians formed two proudly assertive but mutually antagonistic states, India and Pakistan. China's century-long humiliation at the hands of the West exploded in a series of violent revolutions seeking to restore the country to wealth, power, and a place of dignity internationally. "Two Chinas" were eventually established, one communist and the other backed by the United States (Taiwan), the latter of which for two decades spoke for China in the United Nations. The question of China's international representation became one of the great diplomatic issues of the 1950s and '60s. The countries of the Arab world resumed their independence and then insisted, over the objections of their former colonial masters, on exercising full sovereignty throughout their own territories, as Egypt did with respect to the Suez Canal. Anti-imperialist sentiment soon made colonialism unacceptable. By the late 1950s and '60s, new states, mainly in Africa, were being established on an almost monthly basis.

The new states shared the diplomatic forms of the industrialized democracies of the West but not their political culture. Many new states were ill at ease with the values of their former colonial masters and cast about for alternatives drawn from their own history and national experience. Others accepted Western norms but castigated the West for hypocrisy. Envoys began to appear in Western capitals dressed in indigenous regalia to symbolize their assertion of ancient non-Western cultural identities. As they gained a majority at the UN, the newly independent states altered the organization's stance toward colonies, racial issues, and indigenous peoples. Beyond the East-West division of the Cold War, there developed a "North-South" divide between the wealthier former imperial powers of the north and their less-developed former colonies.

The United Nations

The UN, founded in 1946 with only 51 members, was no more successful at healing the North-South rift than it was at healing the East-West one. It was, according to former Indian permanent representative Arthur Lall, "a forum and not a force." On major questions and issues, the UN played only a marginal role, though secretaries-general and their deputies made intense efforts to solve serious but secondary problems such as the resettlement of refugees and persons displaced by war. In the end—as Dag Hammarskjöld, UN secretary-general from 1953 to 1961, remarked—the UN remained only "a complement to the normal diplomatic machinery" of the governments that were its members.

After the larger colonies gained independence, smaller ones followed suit. The new states were often economically underdeveloped and lacked a large educated elite to staff a modern diplomatic corps. The trend continued until even "microstates" of small area and population (*e.g.*, at its independence Nauru had a population under 7,000) became sovereign. The new small states were unable to conduct much diplomacy at first. Many of them accredited ambassadors to only the former colonial power, a key neighbouring state, and the UN.

Over time, the larger of the newly independent states built sizable foreign services modeled on that of the former colonial power or those of the similarly organized services of Brazil and India, which were not complicit in colonialism. The Brazilian foreign ministry and diplomatic service, which were organized and staffed along European lines, enjoyed a reputation as the most professional of such organizations in Latin America. The Indian Foreign Service, which was modeled on the highly respected Indian Administrative Service and initially staffed from its ranks, quickly emerged as a competent practitioner of diplomacy on behalf of a nonaligned, non-Western potential great power. In contrast, the microstates had few missions; instead, they experimented with joint representation, multiple accreditation of one envoy to several capitals, and meeting with foreign envoys in their own capitals. Some states had no foreign ministry and relied on regional powers to represent them.

The exponential growth in the number of states complicated diplomacy by requiring countries—especially the major powers—to staff many different diplomatic missions at once. As state, transnational, and quasi-diplomatic entities proliferated, so did the functions of diplomacy. Leaders met often, but as there was more for diplomats to do, the size of missions of major powers increased.

Regional and subnational organizations. Unlike the UN, some regional organizations achieved significant diplomatic successes. The European Union (EU) effectively promoted trade and cooperation with member states, and the Organization of African Unity (later the African Union) and the Arab League enhanced the international bargaining power of regional groupings of new states by providing a coherent foreign policy and diplomatic strategy. In contrast, the extreme political, economic, and cultural diversity of Asia made it harder to organize effectively; the Organization of American States suffered from the enormous imbalance between the United States and its smaller, poorer, and less-powerful members; and the nonaligned movement was too disparate for long-term cohesion.

Subnational entities also complicated the crowded international scene. Foremost among these was the Palestine Liberation Organization, which had observer status at the UN, membership in the Arab League, and envoys in most of the world's capitals, many with diplomatic status. The African National Congress and the South West Africa People's Organization also conducted long and varied diplomatic campaigns before achieving power in South Africa and Namibia, respectively.

New topics of diplomacy. New topics of diplomacy also abounded, including economic and military aid, commodity-price stabilization, aviation, and allocations of radio frequencies. Career diplomats tended to be generalists, and specialists increasingly came from other agencies as attachés or counselors. Disarmament negotiations, for example, required specialized knowledge beyond the scope of military attachés. Environmental abuse gave rise to a host of topics, such as the law of the sea, global warming, and means of preventing or abating pollution. The complexity of diplomatic missions increased accordingly. By the 1960s, for example, U.S. missions had instituted "country teams," including the ambassador and the heads of all attached missions, who met frequently to unify policy and reporting efforts and to prevent different elements under the ambassador from working at cross-purposes.

Diplomatic personnel and the role of women. In the 1970s the United States, Australia, and some other industrialized democracies (as well as South Africa) broadened recruitment beyond the old elites and attempted to develop foreign services representative of their populations' ethnic diversity. Others, such as Brazil, France, India, and Japan, continued to recruit self-consciously elite services. China and the Soviet Union continued to emphasize political criteria as well as intellectual skills. Overall, however, embassy positions, from the ambassadorial level down, increasingly were filled by professional diplomats. The United States and a handful of other countries, however, continued to appoint wealthy amateurs as ambassadors, treating the most senior diplomatic positions as political spoils.

Although famous female political leaders such as Cleopatra VII, Isabella I, and Elizabeth I were enormously influential in the history of diplomatic relations, historically women largely played a secondary—but substantial—role

as the wives of diplomats. Without large fortunes or many servants, diplomatic wives were forced to shoulder greater burdens as they coped with a nomadic lifestyle, housewifery, hectic social schedules, and endless cooking for obligatory entertaining. The strain became so severe that many ambassadors retired early. In response, Japan adopted the practice of paying diplomatic wives a salary to compensate them for the time they spent entertaining. In 1972 the United States stopped evaluating wives when rating their spouses; entertaining and attending functions were no longer required, though they were still expected. Diplomatic wives also increasingly wished to pursue their own careers. Some of these careers were portable, but when they were not, efforts were made with host countries to find suitable employment.

Women diplomats

In 1923 the Soviet Union became the first country to name a woman as head of a diplomatic mission. The United States, which admitted women to the diplomatic corps beginning in 1925, followed a decade later by appointing a woman as minister to Denmark. France permitted a woman to enter its diplomatic service in 1930, though it still did not appoint women as heads of mission.

After World War II, many women began diplomatic careers, and more women became ambassadors. However, some countries, particularly in the developing world, continued not to hire women as diplomats, and sending women envoys to them was deemed unwise. In 1970 the Vatican rejected a proposed minister from West Germany because she was female. With these exceptions, however, women became an accepted and rapidly growing minority in the diplomatic—including the ambassadorial—ranks. By the end of the 20th century, several American women were serving as ambassadors in Arab and Islāmic countries long considered inhospitable to women.

Before this trend began, women seemed to face almost insuperable difficulties in combining marriage with the nomadic career of diplomacy. Before 1971 the U.S. Foreign Service required women to resign upon marriage. Problems were particularly pronounced for "tandem couples," in which both husband and wife were in the Foreign Service. Because joint postings could not always be arranged, husband and wife often would alternate in taking leave when not posted in adjacent countries. Despite these problems, at the end of the 20th century the U.S. Foreign Service employed some 500 tandem couples.

The end of the Cold War. In 1989, when the Cold War sputtered to a close, there were more than 7,000 diplomatic missions worldwide, most of which were embassies and thus headed by ambassadors. Between World War I and World War II, a few lesser states had been allowed to accredit embassies, but when the United States elevated Latin American missions in the 1940s, a trickle became a flood. Soon legations were the exception, and they disappeared by the last quarter of the 20th century. In addition, numerous international organizations and an array of regional entities, some of them supranational, also now received and sent envoys of ambassadorial rank. For example, some states now accredit three ambassadors to Brussels: to the Belgian government, to the EU, and to NATO.

Meanwhile, the already bewildering variety of tasks assigned to overburdened diplomatic missions continued to grow. Transnational legal issues such as terrorism, organized crime, drug trafficking, international smuggling of immigrants and refugees, and human rights increasingly involved embassies in close liaison with local police and prosecutors. As the 20th century ended, however, the number of diplomatic missions maintained by independent states began to decline because of budgetary constraints, the growing practice within the EU of joint diplomatic representation in capitals of relatively little interest to Europe, and greater willingness to accredit ambassadors simultaneously to several regional states.

Joint diplomatic representation

Even as the number of embassies and diplomats devoted to the conduct of bilateral relations contracted, international organizations and conferences attempting to regulate transnational affairs continued to proliferate. Indeed, the number of nongovernmental entities attempting to influence the work of such organizations and conferences grew even more rapidly: churches, the International Red Cross and similar service and relief organizations, multinational corporations, trade unions, and a host of special-interest groups and professional organizations all developed lobbying efforts aimed at advancing specific transnational agendas. Negotiations over tariffs, debts, and issues of market access meanwhile assumed steadily greater importance and came to involve foreign ministries, ministries of trade, and specialized ambassadors-at-large, as well as resident ambassadors and consular officers.

The end of the Cold War left the foreign relations of many countries without a clear direction. Russia struggled to come to terms with its diminished power and influence, brought about by the political and economic collapse of the former Soviet Union. Deprived of its Soviet enemy and unchallenged as a global power, the United States clung to its alliances but deferred less to its allies and found itself increasingly isolated in international forums. Europe progressed toward greater unity without developing a clear vision of its preferred place in the world. Japan more openly aspired to cast off the restrictions on its international role that its defeat in World War II had imposed, but it did little to define or realize this ambition. China and India, which had seen themselves first as victims of European aggression and then as part of a "Third World" between the American and Soviet-led blocs, began uneasily to emerge as great powers in their own right, in the process reviving elements of their long-forgotten ancient diplomatic doctrines and traditions.

The world map itself changed constantly. As the Soviet Union broke up, the Baltic states resumed their independence, and another wave of new states emerged from the retreat of Russian imperialism from Eastern Europe, the Caucasus, and Central Asia. Meanwhile, various multiethnic states (*e.g.*, Yugoslavia) were torn asunder by rampant nationalism. Ethnic minorities, such as Eritreans, achieved self-determination, and civil strife in countries such as Afghanistan, Rwanda, and Somalia resulted in great suffering and huge refugee flows. The need for international intervention to assist the peoples of failed states seemed to increase constantly. In the course of efforts to assign culpability for large-scale human suffering, the walls of sovereign immunity began to be breached. Even current and former heads of state were no longer exempt from the legal process in both international and national courts. At the beginning of the 21st century, there was a consensus that a transition in the diplomatic order was occurring, though there was disagreement about what kind of new order would emerge.

MODERN DIPLOMATIC PRACTICE

Diplomatic agents. In 1961 the Vienna Convention on Diplomatic Relations replaced the rules that had been adopted in the 19th century. It specified three classes of heads of mission: (1) ambassadors or nuncios accredited to heads of state and other heads of missions of equivalent rank, (2) envoys, ministers, and internuncios accredited to heads of state, and (3) chargés d'affaires accredited to ministers of foreign affairs.

A fourth class that had been established in the 19th century, that of minister-resident, lapsed in the 20th century, during which some variations on the other classes were produced. In 1918 Russia's new regime abolished diplomatic ranks. When the Soviet government gained recognition, it accredited "plenipotentiary representatives." Because they lacked precedence under the rules then prevailing, however, the Soviet Union reverted to the previously used titles. The regime of Muammar al-Qaddafī in Libya sent Peoples' Bureaus, which enjoyed precedence under the current rules. Members of the Commonwealth accredited high commissioners to each other. Finally, the Vatican occasionally sent legates on special missions to Roman Catholic countries and in 1965 began to appoint pro-nuncios. It accredited apostolic nuncios only to those few Roman Catholic states where the papal envoy is always the doyen, or dean, of the diplomatic corps; internuncios elsewhere found themselves in the tiny remaining group of ministers. Hence, the title of pro-nuncio was devised to gain entry into the first class.

Rights and privileges. All heads of mission receive the same privileges and immunities, many of which their aides also enjoy. Diplomatic immunity began when prehistoric rulers first realized that their messengers could not safely convey messages, gather intelligence, or negotiate unless the messengers other rulers sent to them were treated with reciprocal hospitality and dignity. Diplomatic agents and their families are inviolable, not subject to arrest or worse, even in wartime. They are largely outside the criminal and civil law in the host state—even as a witness—though many missions waive some exemptions, especially for parking violations. In the host state, the foreign envoy is free of taxes, and his personal baggage is not inspected by the host state or third states crossed in transit, in which he also enjoys immunity.

Extraterritoriality

The head of mission's residence and the chancellery (usually now called the embassy) are extraterritorial. The legal fiction is maintained that these premises are part of the sending state's territory, not that of the host state; even local firefighters cannot enter "foreign territory" without consent. For this reason, political opponents of harsh regimes often seek asylum in embassies, legations, and nunciatures.

The mission's physical property also enjoys immunities and privileges. The flag and emblem of the sending state may be displayed on the chancellery and on the residence and vehicles of the head of mission. The mission's archives and official correspondence are inviolable even if relations are severed or war is declared. The mission is entitled to secure communication with its government; wireless facilities are either afforded or installed at the mission with the host state's consent. Furthermore, the diplomatic bag and couriers also are inviolable.

In their host country, diplomats enjoy the freedom to articulate their government's policies, even when they are unwelcome to the ears of their hosts. Direct criticism of their host government, its leading figures, or local society may, however, result in a diplomat being asked to leave (*i.e.*, being declared *persona non grata*). By long-standing tradition, host states generally seek to restrain the use of intemperate or insulting language by one country's diplomatic representatives against another's (the so-called "third-country rule").

Credentials. Appointment of a new head of mission is a complex process. To avoid embarrassment, his or her name is informally sounded. If the host country does not object, formal application for *agrément*, or consent, is made by the envoy being replaced. Then the new ambassador is sent forth with a letter of credence addressed by his head of state to the head of the host state to introduce the ambassador as his or her representative. In most major capitals, a copy of credentials is now first provided privately to the foreign minister, after which the new ambassador can deal with the foreign ministry and begin to call on his diplomatic colleagues. Presentation of these credentials to the chief of state is, however, quite formal; in some states with a keen sense of tradition, it may entail riding from the embassy to a palace in an open carriage. The ceremony includes handing over the newly arrived ambassador's letters of credence and those of recall of the predecessor and a short platitudinous speech or brief conversation. The date of the formal presentation of credentials determines an ambassador's order of precedence within the local diplomatic corps. At the UN, credentials are presented without ceremony to the secretary-general. There is no doyen, because turnover is too rapid; instead, the secretary-general annually draws the name of a country from a box, and precedence occurs alphabetically in English beginning with that country.

The appointment of consuls is merely notified; they are entitled to only some diplomatic privileges and immunities. They are located in the major cities of the host country, of which a few may be citizens. Consuls issue visas, but their primary functions are fostering commerce and aiding nationals of the sending state who are in difficulty.

Diplomatic tasks. According to the Vienna Convention, the functions of a diplomatic mission include (1) the representation of the sending state at a level beyond the merely social and ceremonial, (2) the protection within the host state of the interests of the sending state and its nationals, including their property and shares in firms, (3) the negotiation and signing of agreements with the host state when authorized, (4) the reporting and gathering of information by all lawful means on conditions and developments in the host country for the sending government, and (5) the promotion of friendly relations between the two states and the furthering of their economic, commercial, cultural, and scientific relations. Diplomatic missions also perform public services for their nationals, including providing electoral registration, issuing passports and papers for military conscription, referring injured or sick nationals to local physicians and lawyers, and ensuring fair treatment for those charged with or imprisoned for crimes.

Services for citizens and the local public are provided by junior and consular staff, whereas specialized attachés engage in protection and much promotional activity. The ambassador is charged with carrying out all the tasks of the diplomatic mission through subordinates or through personal intervention with local authorities when necessary. The head of mission, the head's spouse, and the deputy spend much time entertaining visiting politicians and attending receptions—at which some business is conducted and information is collected—but representation also entails lodging official or informal protests with the host government or explaining and defending national policy. A diplomat's most-demanding daily activities, however, remain reporting, analyzing, and negotiating.

Reports are filed by telegram, telephone, facsimile, and e-mail, usually on an encrypted basis to protect the confidentiality of information. One of the ambassador's key tasks is to predict a developing crisis, a task accomplished through the gathering of information from an array of sources and the use of experience and expert knowledge in identifying, analyzing, and interpreting emerging key issues and patterns and their implications. The ambassador's duty is to advise and warn, and he is expected to brief his government in detail and without distortion about the content of his conversations with the host foreign minister, the prime minister, and other key officials and politicians.

Beyond these functions, the ambassador negotiates as instructed. Negotiation is a complex process leading to agreement based on compromise, if it reaches agreement at all. The topic of negotiation and the timing of initial overtures are set by the ambassador's foreign ministry. The foreign ministry (perhaps with cabinet involvement) also specifies the diplomatic strategy to be used. Usually this is specific to the goals and circumstances. At the end of World War II, for example, the strategy used by the United States to pursue its goals in Europe was the Marshall Plan, which provided financial assistance to several western and southern European countries. The foreign ministry also establishes broad tactics, often regarding initial demands, bargaining counters, and minimum final position. For the rest, the negotiator, either an ambassador or a special envoy, is in most countries free to employ whatever tactics seem best.

Negotiating tactics

In most negotiations, initial demands far exceed expectations; concessions are as small and as slow as possible, for early concession indicates eagerness and engenders demands for more concessions. There is intermittent testing of the other side's firmness and will for an agreement. There may be indirection, lulling the other party, and bluffing to gain an edge, though it is important for diplomats not to be caught bluffing. Lying in diplomatic negotiations is considered a mistake, but stretching or abridging the truth is permissible. Coercive diplomacy involving the threat of force is risky but cheaper than war; other coercive pressures may include conditions for concessions, such as debt rescheduling. Compensations to sweeten the offer, warnings, and threats speed agreement if well timed, as do deadlines, whether agreed, imposed by external events, or contained in ultimatums.

Negotiations vary according to whether the negotiating states are friends or foes, whether they are of similar or disparate power, whether they genuinely want agreement or are negotiating only for propaganda purposes or to avoid condemnation for refusing to negotiate, and whether their aim is to prolong an existing agreement or to change the

status quo, perhaps redistributing benefits or ending hostilities. Some of the most difficult negotiations plow new ground, as do those that create new cooperative or regulatory institutions, such as the International Sea-Bed Authority, and those that transfer authority, such as the 1984 Sino-British agreement by which Chinese sovereignty over Hong Kong was restored in 1997.

Whatever the problem, the diplomatic negotiator must display reliability and credibility. He must try to create trust and to seem both honest and fair, and he must strive to understand the other side's concerns. Stamina, precision, clarity, courage, patience, and an even temper are necessary, though calculated impatience or anger may be used as a tactic. A skilled negotiator has a sense of timing, knowing when to use threats, warnings, or concessions. Sometimes a third party is discreetly used to facilitate initial contact or to press the sides toward agreement. The negotiator must be persuasive, flexible, tenacious, and creative in devising new solutions or reframing issues from a new angle to convince the other party that agreement is in its interest. Smaller and easier issues are tackled first, building an area of agreement, which is then stressed to create a stake in success, while harder issues are postponed or played down. Through a process of proposal and counterproposal, inducement and pressure, the diplomat keeps talking and, in the last analysis, proceeds by trial and error.

Multilateral negotiations demand the same skills but are more complex. The process is usually protracted and fragmented, with subsidiary negotiations in small groups and occasional cooling-off periods. Skillful representatives of small states often play important roles. Decisions are reached by unanimity, majority, or consensus (to avoid voting). For simplicity, decisions are often made to apply across the board, as with tariff cuts.

Iraq's refusal to end its occupation of Kuwait peacefully in 1990 and the failure of Israel and the Palestinians from the mid-1990s to reach a negotiated settlement of their disputes are sad reminders that when negotiations fail, the consequences can be bloody. In the end, war, not words, remains the ultimate argument of the state. What cannot be decided through dialogue by diplomats or leaders over a negotiating table is often left to be decided on the battlefield or in civil conflict.

Diplomatic agreements. If a negotiation succeeds, the result is embodied in an international instrument, of which there are several types. The most solemn is a treaty, a written agreement between states binding on the parties under international law and analogous to a contract in civil law. Treaties are registered at the UN and may be bilateral or multilateral; international organizations also conclude treaties both with individual states and with each other.

Treaties and conventions

A convention is a multilateral instrument of a law-making, codifying, or regulatory nature. Conventions are usually negotiated under the auspices of international entities or a conference of states. The UN and its agencies negotiate many conventions, as does the Council of Europe. Treaties and conventions require ratification, an executive act of final approval. In democratic countries, parliamentary approval is deemed advisable for important treaties. In the United States, the Senate must consent by a two-thirds vote. Elsewhere, legislative involvement is less drastic but has increased since World War II. In Britain, treaties lie on the table of the House of Commons for 21 days before ratification; other countries have similar requirements. For bilateral treaties, ratifications are exchanged; otherwise, they are deposited in a place named in the text, and the treaty takes effect when the specified number of ratifications have been received.

Agreements are usually bilateral, not multilateral. Less formal and permanent than treaties, they deal with narrow, often technical topics. They are negotiated between governments or government departments, though sometimes nongovernmental entities are involved, as banks are in debt-rescheduling agreements. U.S. presidents have long used executive agreements to preserve secrecy and to circumvent the ratification process.

A protocol prolongs, amends, supplements, or supersedes an existing instrument. It may contain details pertaining to the application of an agreement, an optional arrangement extending an obligatory convention, or a technical instrument as an annex to a general agreement. It may substitute for an agreement or an exchange of notes, which can be used to record a bilateral agreement or its modification.

International instruments have proliferated since World War II: between 1945 and 1965 there were about 2,500 multilateral treaties, more than in the previous 350 years. As the countries of the world have become more interdependent, this trend has continued. Most multilateral agreements are negotiated by conferences. The negotiations are numerous and often protracted.

Conference diplomacy. Professional diplomats are rarely dominant in conferences, where the primary role is usually played by politicians or experts—especially at summits, the most spectacular type. Heads of state or government or foreign ministers meet bilaterally or multilaterally. Summit diplomacy can be risky, a point made in the 15th century by the Burgundian diplomat and chronicler Philippe de Commynes, who wrote, "Two great princes who wish to establish good personal relations should never meet each other face to face, but ought to communicate through good and wise emissaries." Summits also raise expectations; if poorly prepared, they can be disastrous failures. As former U.S. secretary of state Dean Acheson once remarked, "When a chief of state or head of government makes a fumble, the goal line is open behind him." Haste can also lead to bad bargains or murky texts. On the other hand, the development of personal relationships between leaders can be an asset, and political leaders can speed agreement by setting guidelines or deadlines and cutting through bureaucratic thickets. Summits put professional diplomats briefly into the shade but rarely hurt their standing unless there is constant intervention in their work by political leaders or other officials. Indeed, a visit by the foreign minister can be an asset to an ambassador by serving to raise his standing.

Summits

A summit is often preceded or followed by coalition diplomacy. This necessary joint working out of common policies or responses to proposals by cabinet ministers may be fairly informal. Coalitions require cumbersome two-step diplomacy at each stage, arriving at a joint policy and then negotiating with the other party.

Larger conferences are called, often under UN auspices, to address specific problems. The more technical the topic, the larger the role played by specialists. The trend over the last two decades of the 20th century was toward numerous conferences on social, economic, and technical issues. Many conferences produce agreements that create international law, often in new areas. In some cases, the negotiations leading to these agreements are cumbersome. In 1973–75, for example, all 35 states involved in the Geneva Conference on Security and Cooperation in Europe, which led to the Helsinki Accords, participated actively under a unanimity rule. In other cases, the negotiations are protracted, as they were in the Law of the Sea conferences, which lasted more than a decade.

International organizations play several roles in multilateral negotiations, including sponsoring conferences and encouraging coalition diplomacy. The Arab League, the Association of Southeast Asian Nations, and the EU attempt to create a unified policy for their members. Regular meetings of the UN, its agencies, and regional organizations provide forums for parliamentary diplomacy, propaganda, and negotiation. International bureaucracies negotiate with each other and with individual states. This is particularly true of the UN and the EU, the latter of which has assumed some attributes of sovereignty. UN peacekeeping forces have played an important role, and the secretary-general engages in third-party diplomacy to bring feuding states to agreement or at least to keep them talking until the quarrel has faded. States, specialized agencies, and regional entities also conduct third-party diplomacy.

Professional diplomats

Personnel. Diplomatic personnel undergo rigorous selection and training before representing their country abroad. Except in a few cases, those conducting diplomacy are usually professional diplomats or specialists with the title of attaché. Some regimes still use ambassadorships to exile political opponents; others, such as Britain, deviate

from career appointments occasionally for special but nonpolitical reasons. Despite much empirical evidence to suggest that the practice is unwise, U.S. presidents continue to reward major campaign contributors with choice embassies. Even when the ambassador is an amateur, however, other staff members, almost without exception, are career professionals.

Applicants for diplomatic positions generally are university graduates who face grueling oral and written examinations, which few survive. These exams test an applicant's skills in writing, analyzing, and summarizing and the ability to spot essentials and deal with problems, as well as persuasiveness, poise, intelligence, initiative, and stability. As a result of attempts by advanced industrial countries to diversify the educational, ethnic, social, and geographic backgrounds of their diplomatic staffs, foreign-language proficiency is no longer required for entrance into diplomatic training programs; all states educate accepted candidates in languages and etiquette. Despite diversification, the best-educated and most-poised candidates tend to succeed.

All countries agree on the need for proficiency in foreign languages. Diplomats speaking English, French, and Spanish are maintained, and countries often seek candidates with skills in other languages, such as Arabic, Chinese, German, Japanese, Portuguese, Russian. Language training is provided at a foreign service institute, at local universities, or abroad. Most states also stress knowledge of economics, geography, international politics, and law, and many teach their own history and culture. Some provide added academic training; others, including the United States, are more practical in orientation. In the debate over whether career officers should be generalists or specialists, the United States favours modest specialization—for example, in African economics—whereas many states prefer generalists, particularly small countries that cannot afford specialists.

There are three basic approaches to training. Britain and some Commonwealth states couple brief orientation with a long apprenticeship and on-the-job training, some of which occurs in all systems. The French method, also widely imitated, entails intensive training in a school of public administration, in some states with added specialization. India combines the British and French styles in a three-year program. Prospective Brazilian, Egyptian, and German diplomats train for one to three years in an academy that is usually staffed by a combination of senior diplomats and academics and run by the foreign ministry. The United States has no diplomatic academy; instead, it offers highly focused vocational and language training to its diplomats as needed.

Training presents special problems for small new countries, which often use facilities offered by the UN and the Diplomatic Academy of Vienna. A few regional training centres also have been established. Most foreign services, however, rely on a combination of university training and on-the-job apprenticeship.

Structure of the foreign ministry

Once trained, career diplomats serve their foreign ministry abroad or staff it at home. Foreign ministries are similarly organized. They are led by the foreign minister, who is usually a member of the cabinet or dominant political body. In most countries, except those governed by dictatorships, he often belongs to the legislature, though the U.S. secretary of state does not. Some states use the British system of parliamentary undersecretaries to handle legislative responsibilities. Otherwise, except for the minister's staff, employees are civil servants led by a permanent undersecretary or secretary-general, who runs the ministry. The United States is unusual in that it does not have a professional director and the entire top echelon of its diplomatic corps—deputy secretary, undersecretaries and their deputies, and assistant secretaries—is made up of political appointees who are changed with each administration.

Except in the smallest states, foreign ministries are organized both geographically and functionally. The functional departments include administration, personnel, finance, economic affairs, legal affairs, archives, and perhaps offices dealing with science, disarmament, narcotics, and cultural diplomacy. Geographic division is generally by region, subdivided into country desks that deal with accredited embassies and their own missions abroad. Envoys from other states normally see the senior area specialist or the regional assistant secretary, as foreign ministers do not have time to see more than selected ambassadors of a few key countries for especially important questions. Although generalists are preferred in most foreign ministries, some area and country staff will have significant expertise. Despite rotation, this is particularly true in the United States, where career officials specialize in political, economic, administrative, or consular work. All foreign ministries are staffed in varying ratios by two kinds of career diplomat: civil servants based in the capital and foreign service officers on periodic home assignment. Whichever kind they may be and wherever they may serve, they use diplomacy to pursue their country's interests, to engage in international discourse, and to alleviate friction between sovereign states.

BIBLIOGRAPHY. CHAS. W. FREEMAN, JR., *Arts of Power: Statecraft and Diplomacy* (1997), explores the relationship between statecraft and diplomacy and analyzes the functions of diplomats, and *The Diplomat's Dictionary*, rev. ed. (1997), is a compendium of diplomatic quotations and lore. Other important works include JOSÉ CALVET DE MAGALHÃES, *The Pure Concept of Diplomacy* (1988; originally published in Portuguese, 1982); K.M. PANIKKAR, *The Principles and Practice of Diplomacy* (1952, reissued 1957); G.E. DO NASCIMENTO E SILVA, *Diplomacy in International Law* (1972); MARTIN MAYER, *The Diplomats* (1983); R.P. BARSTON, *Modern Diplomacy*, 2nd ed. (1997); and KEITH HAMILTON and RICHARD LANGHORNE, *The Practice of Diplomacy: Its Evolution, Theory, and Administration* (1995).

The history of diplomacy is discussed in RAGNAR NUMELIN, *The Beginnings of Diplomacy: A Sociological Study of Intertribal and International Relations* (1950); FRANK ADCOCK and D.J. MOSLEY, *Diplomacy in Ancient Greece* (1975); RICHARD L. WALKER, *The Multi-State System of Ancient China* (1953, reissued 1971); GANDHI JEE ROY, *Diplomacy in Ancient India* (1981); DONALD E. QUELLER, *The Office of Ambassador in the Middle Ages* (1967); GARRETT MATTINGLY, *Renaissance Diplomacy* (1955, reprinted 1988); WILLIAM JAMES ROOSEN, *The Age of Louis XIV: The Rise of Modern Diplomacy* (1976); HENRY KISSINGER, *A World Restored: Metternich, Castlereagh, and the Problems of Peace, 1812–22* (1957, reissued 1999); IMMANUEL C.Y. HSÜ, *China's Entrance into the Family of Nations: The Diplomatic Phase, 1858–1880* (1960); and ALAN PALMER, *The Chancelleries of Europe* (1983). HENRY KISSINGER, *Diplomacy* (1994), is a critical history of modern statecraft, focusing on the United States. (Sa.M./C.W.F./Ed.)

Disease

Disease is considered to be a harmful deviation from the normal structural or functional state of an organism. A diseased organism commonly exhibits signs or symptoms indicative of its abnormal state. Thus, the normal condition of an organism must be understood in order to recognize the hallmarks of disease. Nevertheless, a sharp demarcation between disease and health is not always apparent.

The study of disease is called pathology. It involves the determination of the cause (etiology) of the disease, the understanding of the mechanisms of its development (pathogenesis), the structural changes associated with the disease process (morphological changes), and the functional consequences of these changes. Correctly identifying the cause of a disease is necessary to identifying the proper course of treatment.

Humans, animals, and plants are all susceptible to diseases of some sort. However, that which disrupts the normal functioning of one type of organism may have no effect on the other types. Thus, the study of disease in each type of organism is treated in a separate section in this article.

For coverage of related topics in the *Macropædia* and *Micropædia,* see the *Propædia,* sections 422 and 423, and the *Index.*

This article is divided into the following sections:

The nature of disease 341
- Major distinctions 341
  - Noncommunicable disease
  - Communicable disease
- Control of disease 344
  - Prevention
  - Treatment

Human disease 345
- Health versus disease 345
- Maintenance of health 346
  - Homeostasis
  - Adaptation
  - Defense against biotic invasion
  - Repair and regeneration
  - Hemostasis
  - Interrelationship of defensive mechanisms
- Disease: signs and symptoms 349
- The causes of disease 350
  - Diseases of genetic origin
  - Chemical and physical injury
  - Diseases of immune origin
  - Diseases of biotic origin
  - Abnormal growth of cells
  - Diseases of metabolic-endocrine origin
  - Diseases of nutrition
  - Diseases of neuropsychiatric origin
  - Diseases of senescence
- Classifications of diseases 360

Diseases of animals 361
- General considerations 361
  - Historical background
  - Importance
  - Role of ecology
- Detection and diagnosis 364
  - Reactions of tissue to disease
  - Methods of examination
  - Tests as diagnostic aids
- Survey of animal diseases 372
  - Infectious and noninfectious diseases
  - Zoonoses
  - Disease prevention, control, and eradication

Diseases of plants 377
- General considerations 377
  - Nature and importance of plant diseases
  - Disease development and transmission
  - Diagnosis of plant diseases
  - Principles of disease control
- Classification of plant diseases by causal agent 384
  - Noninfectious disease-causing agents
  - Infectious disease-causing agents

Bibliography 392

THE NATURE OF DISEASE

Major distinctions

The normal state of an organism represents a condition of delicate physiological balance, or homeostasis, in terms of chemical, physical, and functional processes, maintained by a complex of mechanisms that are not fully understood. In a fundamental sense, therefore, disease represents the consequences of a breakdown of the homeostatic control mechanisms. In some instances the affected mechanisms are clearly indicated, but in most cases a complex of mechanisms is disturbed, initially or sequentially, and precise definition of the pathogenesis of the ensuing disease is elusive. Death in human beings and other mammals, for example, often results directly from heart or lung failure, but the preceding sequence of events may be highly complex, involving disturbances of other organ systems and derangement of other control mechanisms.

The initial cause of the diseased state may lie within the individual organism itself, and the disease is then said to be idiopathic, innate, primary, or "essential." It may result from a course of medical treatment, either as an unavoidable side effect or because the treatment itself was ill-advised; in either case the disease is classed as iatrogenic. Finally, the disease may be caused by some agent external to the organism, such as a chemical that is a toxic agent. In this case the disease is noncommunicable; that is, it affects only the individual organism exposed to it. The external agent may be itself a living organism capable of multiplying within the host and subsequently infecting other organisms; in this case the disease is said to be communicable.

NONCOMMUNICABLE DISEASE

Metabolic defects. Noncommunicable diseases arise from genetically determined metabolic abnormalities present at birth that leave the organism ill-equipped to deal with the natural materials it encounters in its daily life. In human beings, for example, the lack of a certain enzyme necessary for the metabolism of the common amino acid phenylalanine leads to the disease phenylketonuria (or PKU), which appears at a few weeks of age and, if not treated, is often associated with mental retardation. Other metabolic defects may make their appearance only relatively late in life. Examples of this situation are the diseases gout and late-onset, or adult-type, diabetes. Gout results from an accumulation within the tissues of uric acid, an end product of nucleic acid metabolism. Late-onset diabetes results from an impaired release of insulin by the pancreas and a reduction in responsiveness of body tissues to insulin that lead to the inability to metabolize sugars and fats properly. Alternatively, the metabolic fault may be associated with aging and the concomitant deterioration of control mechanisms, as in the loss of calcium from bone in the condition known as osteoporosis. That

these late-developing metabolic diseases also have a genetic basis—that is, that there is an inherited tendency for the development of the metabolic faults involved—seems to be definitely the case in some instances but remains either incompletely understood or uncertain in others.

Deficiency diseases

Environmental hazards. Metabolic derangements also may result from the effects of external environmental factors, a relationship that would be suggested by the apparent confinement of a disease to sharply delimited geographic areas. Notable examples are goitre and mottled enamel of the teeth in humans. The development of goitre is attributable to iodine deficiency in the diet, which leads to compensatory growth of the thyroid gland in a vain effort to overcome the deficiency. The disease tends to occur in inland areas where seafood consumption is minimal and dietary supplementation of iodine—through such items as table salt—does not occur. Mottled enamel of teeth results from consumption of excessive amounts of fluoride, usually in water supplies, but conversely, dental caries (tooth decay) is found to occur to a greater extent in areas in which water supplies are deficient in fluoride. Analogous conditions in herbivorous domesticated animals result from deficiencies in trace elements, such as zinc and selenium, in the soil of pastures and, therefore, also in plants making up the diet. Similarly, plant growth suffers from soil deficiencies of essential elements, particularly nitrogen, potassium, and phosphorus. These conditions can be corrected by adding salts to the diets of domesticated animals and by applying fertilizers to soil.

There also are diseases resulting from toxic substances added to the environment in sufficient amounts to produce symptoms of greater or lesser severity. Although human disorders of this nature are best known, untoward effects of such contamination of the environment occur also in plants and animals. The problems caused by environmental toxic agents are largely, if not entirely, anthropogenic. Best known of the environmental diseases, perhaps, are the occupational diseases, especially those of the respiratory tract, including asbestosis, silicosis, and byssinosis (caused by inhalation of, respectively, asbestos, silica, and cotton dust). Also important in this regard are metal poisoning, as with mercury, lead, and arsenic; poisoning with solvents used in industrial processes; and exposure to ionizing radiation. Of greater importance to the population at large are the diseases that result from exposure to insecticides and atmospheric pollutants. Such diseases usually, though not invariably, are of a chronic nature; they require prolonged exposure to the noxious agent and develop slowly. Environmental diseases of all kinds, however, also may predispose the individual to other diseases; for example, respiratory diseases such as silicosis render the sufferer more susceptible to tuberculosis.

COMMUNICABLE DISEASE

Disease as an ecological problem

Host-parasite relationships. Communicable, or contagious, diseases are those transmitted from one organism to another; infectious diseases are diseases caused in the host by infection with living, and therefore replicating, microorganisms such as animal parasites, bacteria, fungi, or viruses. Practically, these two classes of disease are the same, because infectious diseases generally are communicable, or transmissible, from one host to another, and the causative agent, therefore, is disseminated, directly or indirectly, through the host population. Such spread is an ecological phenomenon, the host serving as the environment in which the parasite lives; complexity arises when the parasite occurs in more than one host species. The host-parasite relationship, therefore, must be considered not only with respect to the individual host-parasite interaction but also in terms of the interrelationship between the host and parasite populations, as well as those of any other host species involved.

Most pathogenic bacteria are obligate parasites; that is, they are found only in association with their hosts. Some, such as staphylococci and streptococci, can proliferate outside the body of the host in nutritive materials infected from host sources. Within the tissues of the host, these organisms set up local infections that spread throughout the body. Still other bacteria, such as the glanders bacillus and the gonococci, meningococci, and pneumococci, are more closely adapted parasites, capable of multiplying outside the body of the host only under the artificial conditions of the laboratory. All these microorganisms have complete cell structures and metabolic capabilities.

A greater degree of dependence on the host is shown by rickettsiae and viruses. Rickettsiae are microorganisms that have the cell structure of bacteria; they exhibit a small degree of metabolic activity outside cells, but they cannot grow in the absence of host tissue. The ultimate in parasitism, however, is that of the viruses, which have no conventional cell structure and consist only of a nucleic acid (either DNA or RNA) wrapped in a protective protein coat. Viruses are obligatory intracellular parasites, capable of multiplying only within the cells of the host, and they have no independent metabolic activity of their own. The genetic information that directs the synthesis of virus materials and certain enzymes enters the host cell, parasitizes its chemical processes, and directs them toward the synthesis of new virus elements.

These various degrees of parasitism suggest that the host-parasite relationship is subject to continuing evolutionary change. The adaptation of the microorganism to its parasitic existence, in this view, is accompanied by progressive loss in metabolic capability, with eventual complete physiological dependence of the parasite on the host.

Parasite specificity. The condition of obligate parasitism is associated with a degree of specificity of the parasite with regard to the host; *i.e.,* the parasite generally is more closely adapted to one species of host than to all others. Microorganisms adapted to plant hosts, with only rare exception, are unable to infect animal hosts, and conversely microorganism parasites of animals rarely occur in plants. A number of host species may be susceptible to infection with a given parasite, and the pattern of host susceptibility need not correspond with taxonomic relationships, including hosts varying as widely as vertebrates and invertebrates.

Adaptation of parasite to host

The ability to produce consistently fatal disease in a host is often of negative survival value to the parasite, because it is quite likely to eliminate quickly all available hosts. Consistent with this, there is a tendency for disease resulting from infection to be less severe when adaptation of the parasite to the host has become close. A change in severity of a disease, presumably resulting from adaptation, has been observed in the case of the spirochete that causes syphilis, with the disease in human beings being less severe today than it was in the 16th century. However, ecological studies of parasitism indicate that it is incorrect to assume that all host-parasite relationships will evolve toward reduced antagonism and that a resultant disease state eventually will be ameliorated. (For further information see BIOSPHERE: *Community ecology.*)

Disease produced in related host species may be either milder or more severe than in the definitive host. In certain cases, adaptation is so close that the parasite is unable to infect any other hosts under natural conditions; this is true of many microorganisms producing disease in humans. On the other hand, natural infection of secondary hosts may occur, leading to severe or fatal disease. Rabies, for example, is a fatal disease in almost all animal hosts. In some species such as the bat, however, the virus may persist for long periods as an asymptomatic infection.

Host resistance. The specificity of pathogenic microorganisms with regard to their hosts is an expression not only of differences in microbial character but also of differing host resistance. The ability of a microorganism to produce disease can be evaluated only in terms of the host reaction, and conversely the resistance, or immunity, of the host can be judged only with regard to its effect on the microorganism. In short, the two are but different facets of the same phenomenon, and either may be evaluated by holding the other constant and varying it. Commonly, for example, virulence of an infective agent is determined experimentally by inoculating groups of hosts with graded doses of the agent and determining, by interpolation, the dose that produces a typical reaction in 50 percent of the host individuals inoculated. This dose is termed the 50-percent-effective dose, or ED_{50}; it is related in inverse

fashion to virulence and in a direct way to resistance. In other words, in a given host, the higher the 50-percent-effective dose, the less virulent the infective organism; or, with a microorganism of known virulence, the higher the ED_{50} with the host it is tested against, the greater the resistance of that particular host. Customarily, in different host species, resistance is expressed as an *n*-fold increase or decrease (with *n* equal to a whole number) in the ED_{50} over that of the normal host species.

Distribution of resistance and virulence

This kind of assay is possible because both virulence and resistance tend to occur in approximately normal, or bell-shaped, frequency distributions; that is, most members of the host and microorganism populations occupy a central position with regard to these properties, exceptional individuals appearing at both extremes. With reference to host resistance, this explains the varied incidence of disease in a host population exposed to a statistically constant dose of the infectious agent. In most practical considerations the dose is only statistically constant, for it varies greatly from one host to another depending on circumstances relating to transfer of the infectious agent. Individual variation in host resistance to infection, however, is due to more than mere numbers of infectious agents encountered; it also results from innate factors in the individual host organism. At any rate, variation in host resistance means that not all individuals making up a population essentially universally susceptible to infection with newly appearing infectious agents will contract the disease on first exposure.

Apparent and inapparent infection. Because infection is not an all-or-nothing affair, individual variation in resistance to disease also results in different degrees of reaction to the infectious agent; *i.e.,* the outcome of the interaction of host and parasite is variable in each individual instance. Some individual hosts show symptoms typical of the disease, and infection is readily recognized. Others, having greater resistance, exhibit symptoms of the disease in only a mild or atypical form, and infection in these individuals may not be clearly recognizable. Still other host organisms become infected with the invading parasite but show no symptoms of the disease. Distinction, therefore, must be made between infection and disease, the former occurring on occasion without any sign of the latter. There may be, of course, no such thing as totally asymptomatic infections. What are taken to be such may be, in fact, only those infections with symptoms occurring beneath the level of observation. Nonetheless, such inapparent infections, or "carrier" states, clearly exist and serve to transmit the infection to susceptible hosts.

The overt consequence of infection of a host population of relatively high resistance is the sporadic occurrence of cases of disease and a high carrier-case ratio. The infection, in other words, is widely prevalent in the host population in asymptomatic form, and the relatively rare observed cases of disease represent the highly susceptible few in the host population making up one extreme of the bell-shaped frequency distribution curve. Examples of human diseases of this kind are poliomyelitis, meningococcal meningitis, and cholera.

This type of irregularity in the occurrence of cases of disease tends to occur in host populations of high, but not too high, resistance to the infectious agent. If host resistance is too high, or too low, the disease will die out: in the former case, because the infective agent is unable to maintain itself and, in the latter, because it eliminates the host. One of the best-known illustrations of the importance of relative host resistance to survival of the parasite is that of the plague bacillus. Plague is primarily a disease of rodents and persists as foci of infection in these hosts. The black rat and the less susceptible gray sewer rat are commonly associated with this disease but are too susceptible to allow its persistence; *i.e.,* the host is destroyed. The infection persists, however, in relatively resistant wild rodents.

Foci of infection

Inheritance of resistance. That there exists genetic control of resistance is suggested by the mere fact of host specificity, and such control has been demonstrated amply by experimental studies on both plant and animal hosts. The former, for example, had wide practical application in the development, by selective breeding, of strains and races of plants of economic importance, especially grains, that are resistant to a wide variety of plant diseases.

In general, resistance developed by selective breeding is only partially specific; that is, the observed resistance to infection with pathogenic microorganisms, and to the toxins of such organisms, is manifested toward groups of related microorganisms producing similar diseases, not to single organisms alone. Although resistance to disease has been found in a few instances to be a function of a single gene, in most cases several genes are involved.

For many years there has been considerable interest in the possibility of differences in resistance to disease associated with the different human populations. While marked differences in morbidity and mortality occur between whites and nonwhites in the United States, for example, it is often difficult to rule out differences in exposure to infection, socioeconomic factors, and differential application of preventive and therapeutic measures in accounting for them. Nevertheless, there are fragmentary indications that there may be sufficient genetic segregation among races to result in differences in resistance to certain diseases. The case fatality rate in tuberculosis appears to be lower in Jews than in others, for example, and gonorrhea seems to be a less serious disease in blacks than in whites.

Racial differences

Epidemiology. The interaction of host and parasite populations constitutes the subject matter of epidemiology (the term being more inclusive than suggested by its relation to the word epidemic). In most instances the epidemiology of infectious disease is characteristic of that disease and is an outgrowth of biological properties of the parasite and the host, including host specificity and the behaviour of the host species as populations.

Aside from the saprophytic microorganisms that occasionally produce disease, most pathogenic microorganisms are adapted sufficiently closely to their hosts that they cannot compete successfully in the physical, chemical, and biological environment outside the host tissues. Exceptions to this generalization occur in the cases of those microorganisms whose life history includes a resistant spore stage. This occurs with various fungi responsible for plant disease, as well as certain parasites of animals. Among the latter are species of the fungus *Coccidioides,* which infect both rodents and humans (producing desert fever in the latter), and the anthrax bacillus, which causes disease in cattle, sheep, and other domesticated animals and occasionally infects humans as well. A disease of animals that can be transmitted to humans is called a zoonosis.

Survival of the parasite ordinarily requires that it be transmitted more or less directly from an infected host organism to a susceptible one. The precise route of infection often is of primary importance, with some microorganisms requiring direct access to internal tissue, some being able to initiate infection on mucous membranes of the nose and throat, and others able to establish primary infections in the intestinal tract. These particular modes of infection generally occur by way of, respectively, biting insects, coughing and sneezing, and contamination of food and water.

The occurrence of a given parasite in more than one host species may markedly affect its epidemiological character; it may persist, for example, in one or another of its hosts as a reservoir of infection, sallying forth to encounter the alternate host only on occasion. It is fairly common to find transmission of a parasite from one vertebrate or plant host to another occurring by means of an insect carrier, or vector. Often animal parasites have intermediate hosts in which one or more phases of their life cycles occur; this results in an obligatory sequence of hosts in the life history of the parasite. With the disease schistosomiasis in humans, for example, the blood flukes responsible for the disease (*Schistosoma* species) spend one phase of their larval life in snails. Under such circumstances, and they are not uncommon, the dissemination of the parasite in a host population is dependent not only on the interaction of the parasite population with that of the host population but also on the interaction of the intermediate or vector host population with both parasite and host populations.

The role of vector hosts

Such interrelationships may be the basis of geographic and seasonal differences in the incidence of disease. An

insect-borne disease transmitted from one host species to another requires the simultaneous presence of all three populations in sufficient numbers for its dissemination. Such a circumstance may be sharply limited by location and season.

Behavioral patterns of host populations often have a great effect on the transmission of infectious agents. Crowding, for example, facilitates the spread of infection. Bovine tuberculosis is largely a disease of domesticated cattle in barns, and the age incidence of the human diseases of childhood is lower in urban than in rural populations, suggesting that in the more crowded urban environment children are exposed to disease at an earlier age.

When a disease is prevalent in an area over long periods of time, it is considered to be endemic in that area. When the prevalence of disease is subject to wide fluctuations in time, it is considered to be epidemic during periods of high prevalence. Epidemics prevailing over wide geographic areas are called pandemics.

Epidemic waves

Epidemic prevalence of disease occurs in a wave, the number of cases rising to a peak and then declining. The period of increase occurs when each case gives rise to more than one additional case—*i.e.,* when the parasite population is growing more rapidly than the host population. The decline occurs when each case gives rise to less than one new case—*i.e.,* when the parasite population begins to die off because it encounters only immune individuals among the host population. The rise and fall in epidemic prevalence of a disease is a probability phenomenon, the probability being that of transfer of an effective dose of the infectious agent from the infected individual to a susceptible one. After an epidemic wave has subsided, the affected host population contains such a small proportion of susceptible individuals that reintroduction of the infection will not result in a new epidemic. Because the parasite population cannot reproduce itself in such a host population, the entire host population is immune to the epidemic disease, a phenomenon termed "herd" immunity.

Following such an epidemic, however, the host population immediately tends to revert to a condition of susceptibility because of (1) the deterioration of individual immunity, (2) the removal of immune individuals by death, and (3) the influx of susceptible individuals by birth. In time the population as a whole again reaches the point at which it is susceptible to epidemic disease. This pattern of rising and falling herd immunity explains why epidemic diseases tend to occur in waves—*i.e.,* exhibit a periodicity in prevalence. The time elapsing between successive epidemic peaks is variable and differs from one disease to another.

Immunity. Humans and all other vertebrates react to the presence of parasites within their tissues by means of immune mechanisms of which there are two types: nonspecific, innate immunity and specific, acquired immunity. Innate immunity, with which an organism is born, involves protective factors, such as interferon, and cells, such as macrophages, granulocytes, and natural killer cells, and its action does not depend on prior exposure to a pathogen. Specific immunity is acquired during the organism's lifetime and involves the activation of white blood cells (B and T lymphocytes), which distinguish and react to foreign substances. B lymphocytes operate by producing antibodies, proteins that neutralize foreign molecules (antigens), while T lymphocytes directly attack invaders. Many immune responses, however, involve both mechanisms.

Although the immune response is primarily defensive in nature, it may contribute in some cases to the pathogenesis of the disease. In rheumatic fever, for example, sensitivity to antigens of the causative streptococcus organism, which cross-react with host tissue antigens, is associated with the progress and adverse aspects of the disease. The immune response to various environmental substances, such as plant pollens and chemotherapeutic drugs, also is responsible for the diseases grouped under the general head of allergies. The immune response itself may become deficient in human diseases involving white cells, such as multiple myeloma, macroglobulinemia, Hodgkin's disease, and chronic lymphocytic leukemia. Such diminished immune responses, however, seem to be of minor significance to the course of these diseases, although, when the disease is sufficiently severe and prolonged, it can increase the risk of opportunistic infections, which can be fatal.

One category of disease is associated with an immune response to antigenic components of the host itself (autoantigens). These diseases, called autoimmune diseases, include rheumatoid arthritis and systemic lupus erythematosus. (For a more detailed explanation of the immunologic system, see IMMUNITY.)

Auto-immune disease

Control of disease

PREVENTION

Most diseases are preventable to a greater or lesser degree, the chief exceptions being the idiopathic diseases, such as the inherited metabolic defects. In the case of those diseases resulting from environmental factors, prevention is a matter of eliminating, or sharply reducing, the responsible material in the environment. Because these materials originate largely from human activities, prevention ought to be a simple matter of the application of well-established principles of industrial hygiene. In practice, however, this is often difficult to achieve.

The infectious diseases may be prevented in one of two general ways: (1) by preventing contact, and therefore transmission of infection, between the susceptible host and the source of infection and (2) by rendering the host unsusceptible, either by selective breeding or by induction of an effective artificial immunity. The nature of the specific preventive measures, and their efficacy, varies from one disease to another.

Quarantine

Quarantine, which is an effective method of preventing transmission of disease in principle, has had only limited success in actual practice. In only a few instances has quarantine achieved prevention of the spread of disease across international borders, and quarantine of individual cases of human disease has long been abandoned as ineffective.

It has not been possible to prevent effectively the dissemination of airborne disease, notably airborne fungal diseases of plants and human diseases of the upper respiratory tract. Nor is disease ordinarily controllable by elimination of reservoirs of infection, such as those that occur in wild animals. There are certain exceptions in which the reservoir of infection can be greatly reduced, however; for example, chemotherapy of human tuberculosis may render individual cases noninfectious, and slaughtering of infected cattle may reduce the incidence of bovine tuberculosis.

When infection is spread less directly, through the agency of living vectors or inanimate vehicles, it is often possible to break one or more of the links connecting the susceptible host with the source of infection. Malaria can be controlled effectively by the elimination of the mosquito vector, and louse-borne typhus in humans can be regulated by disinfestation methods. Similarly, diseases spread in epidemic form through the agency of water or milk are controlled by measures such as the chlorination of public water supplies and the pasteurization of milk.

Artificial immunization

Artificial immunization against certain diseases provides immunity and may be used in these instances, particularly when other methods of control are impractical or ineffective. The mass immunization of children in their early years has been highly effective in the control of diphtheria, smallpox, and poliomyelitis. Under special circumstances, as in certain military populations, it has been possible to control with prophylactic medicinal agents the spread of disease for which effective vaccines have not been developed.

TREATMENT

Treatment of disease in the affected individual is twofold in nature, being directed (1) toward restoration of a normal physiological state and (2) toward removal of the causative agent. The diseased organism itself plays an active part in both respects, having the capacity for tissue proliferation to replace damaged tissue and to surround and wall off the noxious agent, as well as defense and detoxification mechanisms that remove the causative agent and its products

or render them harmless. Therapy of disease supplements and reinforces these natural defense mechanisms.

Metabolic faults also may sometimes be corrected—for example, by the use of insulin in the treatment and control of diabetes—but more often specific therapeutic measures for idiopathic diseases are lacking. However, advances in gene therapy may be able to correct defective genes that result in disease.

When disease is produced by environmental factors, there is commonly no specific treatment; only removal of the affected individual from exposure to the agent generally allows normal detoxification responses to take over. Again, there are notable exceptions, as in the treatment of lead poisoning with ethylenediaminetetraacetic acid, an agent that forms complexes with lead that are excreted by the kidney.

Treatment of infectious diseases is more effective in general; it assumes several different forms. Treatment of diphtheria with antitoxin, for example, neutralizes the toxin formed by the microorganisms, and host defense mechanisms then rid the body of the causative microorganisms. In other diseases, treatment is symptomatic in the sense of restoring normal body function. An outstanding example of this is in cholera, in which disease symptoms result from a massive loss of fluid and salts and from a metabolic acidosis; the highly effective treatment consists of restoring water and salts, the latter including bicarbonates or lactates to combat acidosis. More often, however, therapy is directed against the infecting microorganism by administration of drugs such as sulfonamides or antibiotics. While some of these substances kill the microorganisms, others do not and instead inhibit proliferation of the microorganism and give host defenses an opportunity to function effectively. For other infectious diseases there is no specific therapy. There are, for example, very few antiviral chemotherapeutic agents; treatment of viral diseases is mainly directed toward relief of discomfort and pain, and recovery, if it ensues, is largely a matter of an effective cellular immune response mounted against the invading virus by the host. (W.Bu./D.G.Sc.)

HUMAN DISEASE

Health versus disease

Before human disease can be discussed, the meanings of the terms health, physical fitness, illness, and disease must be considered. Health could be defined theoretically in terms of certain measured values; for example, a person having normal body temperature, pulse and breathing rates, blood pressure, height, weight, acuity of vision, sensitivity of hearing, and other normal measurable characteristics might be termed healthy. But what does normal mean, and how is it established?

The meaning of health

Biological criteria of normality are based on statistical concepts. Body height may be used as an example. If the heights of every individual in a large sample were plotted on a graph, the many points would fall on a bell-shaped curve. At one end of the curve would be the very short people, and at the other extreme the few very tall people. The majority of the points of the sample population would fall on the dome of the bell-shaped curve. At the peak of the dome would be those individuals whose height approaches the average of all the heights. Scientists use curves in determining what they call normal criteria. By accepted statistical criteria, 95 percent of the population measured would be included in the normal range—that is, 47.5 percent above and 47.5 percent below the mean at the very centre of the bell. Looked at in another way, in any given normal biological distribution 5 percent will be considered outside the normal range. Thus the 7-foot (213-centimetre) basketball player would be considered abnormally tall, but that which is abnormal must be distinguished from that which represents disease. The basketball player might be abnormally tall but still have excellent health. Thus, in any statistical analysis of health, the possibility of biological variation must be recognized.

A better example than height of how problems can arise with biological variability is heart size. If the heart is subjected to a greater than normal burden over a long period, it can respond by growing larger (the process is known as hypertrophy). This occurs in certain forms of heart disease, especially in those involving long-standing high blood pressure or structural defects of the heart valves. A large heart, therefore, may be a sign of disease. On the other hand, it is not uncommon for athletes to have large hearts. Continuous strenuous exercise requires a greater output of blood to the tissues, and the heart adapts to this demand by becoming larger. In some cases the decision as to whether an abnormally large heart represents evidence of disease or is simply a biological variant may tax the diagnostic abilities of the physician.

The effects of age

The effects of age introduce yet another difficulty in the attempt to define health in theoretical measured norms. It is well known that muscular strength diminishes in the advanced years of life, the bones become more delicate and more easily fractured, vision and hearing become less sharp, and a variety of other retrogressive changes occur. There is some basis for considering this general deterioration as a disease, but, in view of the fact that it affects virtually everyone, it can be accepted as normal. Theoretical criteria for health, then, would have to be set for virtually every year of life. Thus, one would have to say that it is normal for a man of 80 to be breathless after climbing two flights of stairs, while such breathlessness would be distinctly abnormal in an agile child of 10 years of age. Moreover, an individual's general level of physical activity significantly alters his ability to respond to the ordinary demands of daily life. The amount of muscular strength possessed by an 80-year-old man who has remained physically active would be considerably more than that of his fragile friend who has led a confined life because of his dislike of activity. There are, therefore, many difficulties in establishing criteria for health in terms of absolute values.

Health might be defined better as the ability to function effectively in complete harmony with one's environment. Implied in such a definition is the capability of meeting—physically, emotionally, and mentally—the ordinary stresses of life. In this definition health is interpreted in terms of the individual's environment. Health to the construction worker would have a dimension different from health to the bookkeeper. The healthy construction worker expects to be able to do manual labour all day, while the bookkeeper, although perfectly capable of performing sedentary work, would be totally incapable of such heavy labour and indeed might collapse from the physical strain; yet both individuals might be termed completely healthy in terms of their own way of life.

Physical fitness

The term physical fitness, although frequently used, is also exceedingly difficult to define. In general it refers to the state of optimal maintenance of muscular strength, proper function of the internal organs, and youthful vigour. The champion athlete prepared to cope not only with the commonplace stresses of life but also with the unusual illustrates the concept of physical fitness. To be in good physical condition is to have the ability to swim a mile to save one's life or to slog home through snowdrifts when a car breaks down in a storm. Some experts in fitness insist that the state of health requires that the individual be in prime physical condition. They prefer to divide the spectrum of health and disease into (1) health, (2) absence of disease, and (3) disease. In their view, those who are not in prime condition and are not physically fit cannot be considered as healthy merely because they have no disease.

Health involves more than physical fitness, since it also implies mental and emotional well-being. Should the angry, frustrated, emotionally unstable person in excellent physical condition be called healthy? Certainly this individual could not be characterized as effectively functioning in complete harmony with the environment. Indeed, such

an individual is incapable of good judgment and rational response. Health, then, is not merely the absence of illness or disease but involves the ability to function in harmony with one's environment and to meet the usual and sometimes unusual demands of daily life.

The definitions of illness and disease are equally difficult problems. Despite the fact that these terms are often used interchangeably, illness is not to be equated with disease. A person may have a disease for many years without even being aware of its presence. Although diseased, this person is not ill. Similarly, a person with diabetes who has received adequate insulin treatment is not ill. An individual who has cancer is often totally unaware of having the disorder and is not ill until after many years of growth of the tumour, during which time it has caused no symptoms. The term illness implies discomfort or inability to function optimally. Hence it is a subjective state of lack of well-being produced by disease. Regrettably, many diseases escape detection and possible cure because they remain symptomless for long years before they produce discomfort or impair function.

Illness and disease

Disease, which can be defined at the simplest level as any deviation from normal form and function, may either be associated with illness or be latent. In the latter circumstance, the disease will either become apparent at some later time or will render the individual more susceptible to illness. The person who fractures an ankle has an injury—a disease—producing immediate illness. Both form and function have been impaired. The illness occurred at the instant of the development of the injury or disease. The child who is infected with measles, on the other hand, does not become ill until approximately 10 days after exposure (the incubation period). During this incubation period the child is not ill but has a viral infectious disease that is incubating and will soon produce discomfort and illness. Some diseases render a person more susceptible to illness only when the person is under stress. Some diseases may consist of only extremely subtle defects in cells that render the cells more susceptible to injury in certain situations. The blood disease known as sickle cell anemia, for example, results from a hereditary abnormality in the production of the red oxygen-carrying pigment (hemoglobin) of the red cells of the blood. The child of a mother and father who both have sickle cell anemia will probably inherit an overt form of sickle cell anemia and will have the same disease as the parents. If only one parent has sickle cell anemia, however, the child may inherit only a tendency to sickle cell anemia. This tendency is referred to by physicians as the sickle cell trait. Individuals having such a trait are not anemic but have a greater likelihood of developing such a disease. When they climb a mountain and are exposed to lower levels of oxygen in the air, red blood cells are destroyed and anemia develops. This can serve as an example of a disease or a disease trait that renders the affected person more susceptible to illness.

Disease, defined as any deviation from normal form and function, may be trivial if the deviation is minimal. A minor skin infection might be considered trivial, for example. On the eyelid, however, such an infection could produce considerable discomfort or illness. Any departure from the state of health, then, is a disease, whether health be measured in the theoretical terms of normal measured values or in the more pragmatic terms of ability to function effectively in harmony with one's environment.

Maintenance of health

Health is not a static condition but represents a fluid range of physical and emotional well-being continually subjected to internal and external challenges such as worry, overwork, varying external temperatures, mechanical stresses, and infectious agents. These constantly changing conditions require the adjustment of the function of the various systems within the body. Mechanisms are continually at work to maintain a constant internal environment called by the French scientist Claude Bernard the *milieu intérieur.* The maintenance of this relatively constant internal environment is known as homeostasis. On a hot summer day, for example, the body is challenged to maintain its normal temperature of 98.6° F (37° C). Sweating represents a mechanism by which the skin is kept moist. By the evaporation of the moisture, heat is lost more rapidly. The hot day, therefore, represents a challenge to homeostasis. On a cold day gooseflesh may develop, an example of a homeostatic response that is a throwback to mechanisms in lower animals.

Definition of homeostasis

Bacteria, viruses, and other microbiological agents are obvious challenges to health. The body is able, to a considerable extent, to protect itself and adjust to challenges, and, to the extent that it is successful, the state of health is maintained. While health is often thought of as fragile and subject to many onslaughts, it is, in fact, a ruggedly guarded state protected by a host of highly efficient internal mechanisms.

Some of the mechanisms vital to the maintenance of health include (1) the maintenance of the internal environment, or homeostasis, (2) adaptation to stress situations, (3) defense against microbiological agents, such as bacteria and viruses, (4) repair and regeneration of damaged tissue or cells, and (5) clotting of the blood to prevent excessive bleeding. Each of these areas will be discussed briefly. Despite these separate considerations, the commonality of purpose—the preservation and maintenance of health—must not be lost sight of. Insofar as each of these mechanisms works to maintain a constant internal environment, it can be considered as a homeostatic mechanism. Later, when disease is discussed, it will be apparent that to a considerable extent disease represents a failure of homeostasis and the other defensive responses listed above.

HOMEOSTASIS

As noted earlier, the term homeostasis refers to the maintenance of the internal environment of the body within narrow and rigidly controlled limits. The major functions important in the maintenance of homeostasis are fluid and electrolyte balance, acid-base regulation, thermoregulation, and metabolic control.

Fluid and electrolyte balance. This term refers to the controlled partition of water and major chemical constituents among the cells and the extracellular fluids of the body. The human body is basically a collection of cells grouped together into organ systems and bathed in fluids, most notably the blood. The intracellular fluid is the fluid contained within cells. The extracellular fluid—the fluid outside the cells—is divided into that found within the blood and that found outside the blood; the latter fluid is known as the interstitial fluid. These fluids are not simply water but contain varying amounts of solutes (electrolytes and other bioactive molecules). An electrolyte (sodium chloride, for example) is defined as any molecule that in solution separates into its ionic components and is capable of conducting an electric current. Cations are electrolytes that migrate toward the negative pole of an electric field; anions migrate toward the positive pole. The electrolyte composition of the various fluid compartments is summarized in Table 1.

Intracellular and extracellular fluid

It is apparent from this table that the ionic compositions of the intracellular and extracellular fluids are significantly different. The major cation of extracellular fluid is sodium.

Table 1: Principal Electrolytes of the Body Fluids

electrolyte	extracellular fluid*	intracellular fluid†
Cations (+ electrical charge)		
Sodium (Na^+)	142 mEq/l‡	10 mEq/l
Potassium (K^+)	4 mEq/l	160 mEq/l
Calcium (Ca^{++})	5 mEq/l	—
Magnesium (Mg^{++})	3 mEq/l	35 mEq/l
	154 mEq/l	205 mEq/l
Anions (− electrical charge)		
Chloride (Cl^-)	103 mEq/l	2 mEq/l
Bicarbonate (HCO_3^-)	27 mEq/l	8 mEq/l
Phosphate (PO_4^{3-})	2 mEq/l	140 mEq/l
Sulfate (SO_4^{2-})	1 mEq/l	—
Protein	16 mEq/l	55 mEq/l
Organic acid	5 mEq/l	—
	154 mEq/l	205 mEq/l

*Approximate values in the blood plasma. †Approximate values for the muscle cells. ‡mEq/l = milliequivalents per litre.

The major anion of the extracellular fluid is chloride, while bicarbonate is the second most important. In contrast, the major cation of the intracellular fluid is potassium, and the major anions are proteins and organic phosphates. The marked differences in sodium and potassium concentrations between the intracellular and extracellular fluid of cells are not fortuitous but are due to active transport by energy-dependent ion pumps located in cell membranes. The pumps continuously move sodium ions out of the cell and potassium ions into the cell. The intracellular and extracellular compartments are thus closely integrated and interdependent: changes in one have immediate effects on the other. In clinical medicine most measurements of electrolyte concentration are performed on the extracellular fluid compartment, notably the blood serum. The values given in Table 1 remain fairly constant on a day-to-day basis, in spite of various dietary intakes of food and water.

It is the primary task of the kidneys to regulate the various ionic concentrations of the body. Any abnormality in these concentrations can produce serious disease; for instance, the normal sodium concentration in the serum (the blood minus its cells and clotting factors) ranges from 136 to 142 milliequivalents per litre, while the normal potassium level in the serum is kept within the narrow range of 3.5 to 5 milliequivalents per litre. A rise in the serum potassium to perhaps 6.2 milliequivalents per litre, as can occur when large numbers of cells are severely injured or die and potassium ions are released, could cause serious abnormalities in the performance of the heart by disturbing the regularity of the nervous impulses that maintain the heart's rhythm.

The state of hydration

The total amount of body water is also maintained at fairly constant levels from day to day by the combined action of the central nervous system and the kidneys. If one were to refrain from drinking any water for a few days, the thirst centre, located in the hypothalamus deep within the brain, would send out messages that would be translated into the feeling of thirst. At the same time a hormone from the posterior pituitary gland known as antidiuretic hormone (ADH; vasopressin) would be secreted. This hormone, released into the bloodstream, would reach the kidneys, where it would signal the kidneys to retain water and not excrete it. Should too much water be ingested, ADH secretion would be turned off, and the kidneys would promptly excrete the excess amount.

Acid-base equilibrium. The acidity of the body fluids is maintained within narrow limits. This acidity is expressed in terms of the pH of a solution, values exceeding 7 representing alkalinity and less than 7 acidity. The pH of a solution is an expression of the amount of hydrogen ion present. Increases in hydrogen ion concentration cause a lowering of the pH, and, conversely, decreases in the hydrogen ion concentration raise the pH. Any abnormal process raising the hydrogen ion concentration in the body fluids produces a state of disease referred to as acidosis; one that causes the concentration to be lowered results in alkalosis.

In health the blood is slightly alkaline, being kept at a pH of 7.35 to 7.45, a narrow range which must be maintained for the optimum operation of the many chemical reactions that go on constantly in the body. Alterations in the blood pH occur in many diseases, particularly of the lungs and kidneys, organs whose functions include regulation of the body pH.

Thermoregulation. As has been said above, the temperature of the body is kept nearly constant at 98.6° F (37° C). Fluctuations within a few tenths of a degree are perfectly compatible with health. Wider swings in temperature are usually indicative of disease, and thus body temperature is an important factor in assessing health. Body temperature is regulated by a thermostatic control centre in the hypothalamus. A rise in body temperature initiates a chain of events leading to an increase in the rate of breathing and in sweating, two processes that serve to lower the body temperature. Similarly, a decrease in body temperature, perhaps occasioned by a chilly winter walk, leads to increased heat-producing activity such as the muscular contractions of shivering—again mediated by the thermostatic control centre in the hypothalamus.

Metabolic control. In essence, metabolism involves all the physical and chemical processes by which cells are produced and maintained. Included under this broad umbrella are the regulation of fluid and electrolytes, the maintenance of plasma protein levels adequate for the building and repair of cells, and control of the amounts of sugar (glucose) and fats (lipids) in the blood so as to provide sufficient amounts for all the energy-producing activities of the cells. (The main treatment of this subject is contained in the article METABOLISM.)

Blood glucose level

The control of blood glucose levels is a good example of homeostasis. Most of the glucose utilized by the body is derived from the dietary intake of various forms of sugars and starches. These are digested within the intestinal tract into the simplest forms of carbohydrate (monosaccharides). Glucose, galactose, and fructose are the principal monosaccharides. These are absorbed from the intestines into the blood and enter the liver. Here all are eventually converted to glucose. The glucose may be utilized by the liver cells in part as a source of readily available energy, or it may be polymerized and stored as glycogen, but most of it enters the general circulation of the body and contributes to the blood glucose level. Blood glucose may also be derived in times of need by the conversion of the stored glycogen into glucose.

When food is eaten, there is a temporary rise in the blood glucose level known as alimentary hyperglycemia (high blood glucose level). Mechanisms are activated that stimulate the pancreas to secrete the hormone insulin. This hormone makes it possible for cells to utilize the glucose by facilitating its transport (carriage) across the membranes of cells into their interior, where it can enter the complex chemical reactions that provide the cell with energy. By virtue of insulin secretion, the cells receive adequate amounts of glucose, and the blood glucose levels are returned to the normal range, somewhere between 70 and 110 milligrams per 100 millilitres of blood.

ADAPTATION

Adaptation refers to the ability of cells to adjust to severe stresses and achieve altered states of equilibrium while preserving a healthy state. In the human body the large bulging muscles of an individual engaged in heavy labour are a good example of cellular adaptation. Because of the heavy demand for work from these muscles, each of the individual muscle cells within the labourer's arms and legs becomes larger (hypertrophic). This enlargement is caused by the formation of increased numbers of tiny fibres (myofilaments) that provide the contractile power of muscles. Thus, while the normal muscle cell might have 2,000 myofilaments, the hypertrophied cell might have 4,000 myofilaments. The workload can now be divided evenly among twice as many myofilaments, and the muscle cell is capable of more work. The cells are completely normal and, in fact, are more robust than their fragile cousins. The individual can do heavy work all day without excessive fatigue, and no cell injury results from the heavy workload. A new level of equilibrium has been achieved by the process of cellular hypertrophy. A person with this type of muscular development can be considered to be in excellent physical condition, capable of meeting emergency situations such as running from a fire or catching a train without the dangers that might be encountered by a person who has not undergone such a development.

The enlarged muscle cell

Inhabitants of high altitudes adapt to the lowered amounts of oxygen within the air by developing an increased number of red blood cells (a condition called secondary polycythemia). The greater number of red cells in the blood are capable of absorbing more oxygen from the air breathed into the lungs, and thus the person who lives in high altitudes makes better use of the slender oxygen content of the air.

Thus, adaptation is a mechanism by which the body preserves and maintains its health by adjusting to alterations in the conditions under which it functions.

DEFENSE AGAINST BIOTIC INVASION

Human beings are surrounded by a microscopic menagerie of organisms, most of which pose no threat and some

Four levels of defense

of which are beneficial. Organisms capable of producing disease are pathogens. The maintenance of health requires defense against biotic invasion. There are four levels of defense in the body: (1) the intact skin and linings of the various orifices of the body (such as the mouth, nose, throat), (2) a widely dispersed system of cells capable of destroying invaders, (3) the capability of mounting an inflammatory reaction that destroys offenders, and (4) the capability of developing an immune response that helps to bring about further neutralization and destroy any attackers.

Maintenance of the integrity of skin and mucosal linings. With rare exception, pathogenic organisms cannot penetrate the intact covering and linings of the body. Indeed, if one were to take samples of the bacteria found on the skin, one would find large numbers of potentially harmful organisms that represent no threat unless the skin is punctured or the linings of the body are in some way injured. There are exceptions to this generalization, and a few biotic agents probably can penetrate intact mucosal surfaces. The bacterium *Salmonella typhi* that causes typhoid fever is thought to penetrate the normal lining of the gastrointestinal tract. Nevertheless, the intact skin and mucosal linings are primary protective barriers in the maintenance of health. The skin serves as a barrier to the external world, and the mucus-secreting and ciliated membranes of the upper respiratory tract trap inhaled foreign material and bacteria, transporting them to the pharynx where they are either swallowed or expelled by coughing. Potentially harmful bacteria can be introduced into a cut, which thus provides a portal of entry for organisms that may then cause an infection. By adequate washing, at least sufficient numbers of bacteria are flushed out to prevent the infection. Irritation of the skin from any cause or irritation of the throat by habitual smoking of tobacco impairs the integrity of these barriers and predisposes the area to invasion by potentially harmful organisms. The body has ingeniously contrived to place further roadblocks in the way of invaders. The saliva and the secretions in the stomach, for example, contain enzymes and acids that also destroy most organisms. Thus, humans have an effective enclosing barrier that provides protection against biotic attack.

Phagocytic cells of the body. Phagocytosis is the process by which certain cells ingest particulate material. When a phagocytic cell comes in contact with some particle such as a bacterium or even inert material such as dust, the cytoplasm of the cell (the cell substance outside its nucleus) flows around the object and forms a phagocytic vesicle. The phagocytic vesicle containing the particle then fuses with a lysosome (a membrane-enclosed sac that contains digestive enzymes). If the chemical composition of the foreign substance permits its degradation by the enzymes, it is destroyed. If the ingested material is resistant to digestion, it is retained within the phagocyte and is thus effectively removed from further interaction with the host. Phagocytic cells abound in the body; they serve as a second line of defense against most biotic invasion.

There are two groups of phagocytic cells, white blood cells—polymorphonuclear leukocytes—and tissue cells. The white blood cells are able to migrate through blood-vessel walls in areas of inflammation or infection, where they may phagocytize foreign material such as bacteria. Moreover, in inflammatory and infectious states, the total number of white cells in the body increases (leukocytosis). Thus the population of phagocytic cells is expanded when the cells are needed in the body's defense.

The second group of phagocytes consists of cells that are usually firmly fixed within tissues and are known as the reticuloendothelial system. The cells in this system are designated by a variety of names depending on their location (*e.g.*, Kupffer cells in the liver, macrophages or histiocytes in loose connective tissue). They are particularly abundant in the spleen, liver, lymph nodes, and bone marrow but are also scattered throughout the blood vessels and virtually all the other tissues of the body. If, for example, bacteria do find a portal of entry but the bacterial invasion is not too massive and the organisms are not too virulent, these phagocytic cells are capable of engulfing and destroying them before they can cause injury.

The inflammatory response. Whenever cells are damaged or destroyed, a series of vascular and cellular events known as the inflammatory response is set in motion. This response is protective of health in that it destroys or walls off injurious influences and paves the way for the restoration of normality. The sequence of events is as follows: in an area of injury (as in a bacterial infection), cells release substances that cause the small blood vessels in the affected area to become dilated (vasodilation) and thus increase the blood flow to the injured area; at the same time, clear fluid leaks out of the vessels into the area; this fluid tends to dilute any harmful substances in the area of injury; next, white cells from the blood flow out of the blood vessels into the damaged area and phagocytize the bacteria and dead cells; the resulting mixture of dead cellular debris and white blood cells is known as pus.

Major signs of inflammation

The major signs of inflammation are redness and increased heat (caused by blood-vessel dilation), swelling (resulting from the accumulation of fluid), and pain. The last of these is one of the cardinal signs of all inflammatory responses. Pain in inflammation is caused by substances released by damaged tissues that render local nerve endings more sensitive to stimulation. Inflammation can be classified as either acute or chronic. Acute inflammation, such as may be seen around a skin cut, lasts for only a few days and is characterized microscopically by the presence of polymorphonuclear leukocytes. Chronic inflammation is of longer duration and is characterized microscopically by the presence of lymphocytes, monocytes, and plasma cells and, in general, is associated with little fluid exudation.

Because of the pain and swelling, the inflammatory response is often viewed as an unwelcome event following injury. Yet it is important to recognize that it is the first step in the healing process and represents an important protective response in the maintenance of health.

The immune response. The immune reaction is one of the most important defense mechanisms against biotic invasion and is therefore vital to the preservation of health. The devastating effects of acquired immune deficiency syndrome (AIDS) and other conditions that suppress or destroy the immune system are cases in point (see below *The causes of disease: Diseases of immune origin*).

The immune response is a relatively recent evolutionary development found only in vertebrates. This complex system has multiple components, which include antigens, antibodies, complement, and various types of white blood cells such as B and T lymphocytes. The interaction of these components collectively results in a reaction that serves to protect the host from the potentially adverse effects of infectious organisms. Antigens are proteins, polysaccharides (complex carbohydrates), or foreign substances that trigger an immune response; they include molecules that are important constituents of bacteria, viruses, and fungi and substances that mark the surfaces of foreign materials such as pollen or transplanted tissue. Antibodies, or immunoglobulins, are proteins raised against specific antigens; they are formed in the lymph nodes and bone marrow by mature B lymphocytes called plasma cells and are released into circulation to bind and neutralize antigens located throughout the body. This type of response, called humoral immunity, is active mainly against toxins and free pathogens (those not ingested by phagocytes) in body fluids. A second type of response, called cell-mediated immunity, does not yield antibodies but instead generates T lymphocytes that are reactive against specific antigens. This defense is exhibited against bacteria and viruses that have been taken up by the host's cell as well as against fungi, transplanted tissue, and cancer cells. In each case the immune response prevents the invaders from causing further damage to the host. The complement system is a group of proteins found in the blood that facilitates the immune response by both attracting phagocytes to the area of invasion and forming a complex that results in lysis of the foreign cell.

Action of complement

Two remarkable qualities of the immune system are specificity and memory. When an antigen enters the body, it elicits production of either a specific antibody or specific immunologically competent cells; that is, the antibody or the cells will neutralize only the antigen that evokes them.

Furthermore, the system exhibits what appears to be memory: once challenged by an antigen, such as the measles virus, the body "remembers" it for years and perhaps for life. The child who has an attack of measles becomes immune for life. If the child is exposed to this specific antigen at a later date, the immune system recognizes it and responds and thereby prevents a reinfection. Indeed, these two characteristics of the immune system, specificity and memory, serve as the basis for preventive immunization. Inoculation of infants or children with inactivated or attenuated biotic agents will cause the immune system to be made alert to such an antigen should it appear at a later date. Poliomyelitis, for example, once dreaded as a cause of paralysis and death, has been effectively controlled if not abolished with the polio vaccine.

Thus, the immune system is a vital part of the defense against biotic invasion. However, if it malfunctions, the immune system may also cause disease.

REPAIR AND REGENERATION

By replacing damaged or destroyed cells with healthy new cells, the processes of repair and regeneration work to restore an individual's health after injury. Unlike the salamander, which is capable of regenerating a limb if it is lost, human beings cannot regenerate whole organs or limbs. If one kidney is destroyed by disease, it is permanently lost. However, the remaining contralateral kidney, if normal, is capable of limited regeneration to compensate for the decrease in kidney mass. The many cell types of the body have varying capacities for regeneration.

Regeneration

Regeneration is the production of new cells exactly like those destroyed. Of the three categories of human cells—(1) the labile cells, which multiply throughout life, (2) the stable cells, which do not multiply continuously but can do so when necessary, and (3) the permanent cells, incapable of multiplication in the adult—only the permanent cells are incapable of regeneration. These are the brain cells and the cells of the skeletal and heart muscles.

Labile cells are those of the bone marrow, the lymphoid tissues, the skin, and the linings of most ducts and hollow organs of the body.

Stable cells are found in the liver, in many of the glands of the body, such as the pancreas and salivary glands, in the lining of the kidney tubules, and in the connective tissues. Normally these cells do not divide unless some are destroyed by disease or injury and must be replaced.

If only a small area of the liver (made up of stable cells) is damaged or destroyed, unaffected cells around the area of injury can replace those that were lost. When large areas of the liver are destroyed, however, cellular regeneration cannot occur, and the area of cell loss is replaced by new healthy connective-tissue cells, which produce scars. If a heart attack occurs, a certain number of heart muscle cells (permanent cells) are killed because of loss of blood supply. Because heart muscle cells cannot regenerate, the area of injury is replaced by a scar (if the patient survives). Such repair is by no means perfect, but it nonetheless permits restoration of reasonable heart function with perhaps only a slightly reduced level of health, depending on the number of heart muscle cells that have been lost.

Cellular regeneration in humans is limited by many other factors, such as the availability of blood supply and a supporting connective tissue. When the blood vessels and supporting cells (connective tissue) are destroyed in the liver along with the liver cells, perfect reconstitution of the liver is not possible. There may be some regrowth of liver cells, but they do not form the normal liver architecture, and the newly regenerated cells cannot function because they do not have an appropriate orientation to the blood vessels and bile ducts.

HEMOSTASIS

Another mechanism of defense is hemostasis, the prevention of loss of blood from damaged blood vessels by formation of a clot. (This process is covered more at length in the article BLOOD: *Bleeding and blood clotting*.) Simply stated, a break in a blood vessel leads to activation of a complex sequence of events that results in the formation of a solid plug of platelets, red blood cells, and fibrin (a fibrous protein formed from fibrinogen). This plug, or clot, seals the damaged vessel and prevents further loss of blood (hemorrhage). The numerous components of the blood called clotting factors contribute in sequential fashion to the formation of the clot. (The clotting factors are commonly referred to by a roman numeral rather than by name. Fibrinogen, for example, is clotting factor I.) A defect in one of these factors can undermine hemostasis; for example, the absence of clotting factor VIII leads to hemophilia A, a disorder of uncontrolled bleeding.

INTERRELATIONSHIP OF DEFENSIVE MECHANISMS

The homeostatic and defensive mechanisms involved in maintaining a constant internal environment are complex and yet wonderfully coordinated. Thus, the normal state of health is not a static condition but exists rather within a narrow range maintained by the coordinated responses of many systems and mechanisms. Health requires the proper function of all these controls. Disease may begin in a single organ or system, but the interdependence and close coordination of the many bodily functions, which cooperate so beautifully in health, may be upset by a chain reaction when one breaks down. A disease of the kidney leading to abnormal retention of sodium, for example, can cause hypertension (high blood pressure). Prolonged hypertension in turn can induce heart failure, and this can result in the abnormal collection of fluid in the lungs. The impairment of respiratory function may then result in a sudden rise in the level of carbon dioxide in the blood, which brings with it further complications. Similarly, if the normal inflammatory response malfunctions, a trivial skin infection (popularly known as a pimple) can enlarge into a boil (a furuncle). The responsible bacterial agents may proliferate in the local site and penetrate small blood vessels to seed the bloodstream, thus causing a generalized infection (septicemia or bacteremia). Such a widespread infection is extremely serious and may cause secondary infections of the heart (endocarditis) or of the coverings of the brain (meningitis) and end in death of the host.

Chain reactions in disease

Thus, health implies the proper functioning of the homeostatic mechanisms that have just been described, including those systems involved in the defense of health. The state of disease basically represents a failure of these mechanisms. Although one tends to think of disease in terms of offending agents, these agents are able to produce disease only by their ability to disrupt normal homeostasis, and it is precisely those disruptions that are the manifestations of disease.

Disease: signs and symptoms

Disease may be acute, chronic, malignant, or benign. Of these terms, chronic and acute have to do with the duration of a disease, malignant and benign with its potentiality for causing death.

An acute disease process usually begins abruptly and is over soon. Acute appendicitis, for example, is characterized by the sudden onset of nausea, vomiting, and pain usually localized in the lower right side of the abdomen. It usually requires immediate surgical treatment. The term chronic refers to a process that often begins very gradually and then persists over a long period. For example, ulcerative colitis—an inflammatory condition of unknown cause that is limited to the colon—is a chronic disease. The disease is characterized by relapsing attacks of bloody diarrhea that persist for weeks to months. These attacks alternate with asymptomatic periods that can last from weeks to years.

The terms benign and malignant, most often used to describe tumours, can be used in a more general sense. Benign diseases are generally without complications, and a good prognosis (outcome) is usual. A wart on the skin is a benign tumour caused by a virus; it produces no illness and usually disappears spontaneously if given enough time (often many years). Malignancy implies a process that, if left alone, will result in fatal illness. Cancer is the general term for all malignant tumours.

Diseases usually are indicated by signs and symptoms. A sign is defined as an objective manifestation of disease that

can be determined by a physician; a symptom is subjective evidence of disease reported by the patient. Each disease entity has a constellation of signs and symptoms more or less uniquely its own; individual signs such as fever, however, may be found in a great number of diseases. Some of the common manifestations of disease—as they relate to an imbalance of normal homeostasis—are taken up in this section. They are covered more at length in the article DIAGNOSIS AND THERAPEUTICS.

Fever is an abnormal rise in body temperature. It is most often a sign of infection but can be present whenever there is tissue destruction, as, for example, from a severe burn or when large amounts of tissue have died because of lack of blood supply. Body temperature is controlled by the thermostatic centre in the hypothalamus. Certain protein and polysaccharide substances called pyrogens, released either from bacteria or viruses or from destroyed cells of the body, are capable of raising the thermostat and causing a rise in body temperature. Fever is a highly significant indicator of disease.

Increased number of white blood cells

An increase in the number of circulating phagocytic white blood cells (leukocytosis), mentioned above (see *Maintenance of health: Defense against biotic invasion: Phagocytic cells of the body*), is one of the more common manifestations of disease. The stimulus for such an event may be any inflammatory process in the body, such as is caused by bacteria, viruses, or any process that leads to the destruction of cells. Such leukocytosis is reflected in the white blood cell count, which may be substantially elevated above the normal upper value of 10,000 cells per cubic millimetre of blood.

The pulse rate is another easily obtainable and important piece of information. The heart rate varies with the level of physical activity: the heart beats faster during exercise and more slowly during rest. An inappropriate heart rate (or pulse) may be indicative of disease. The heart rate increases in the feverish patient. A weak, rapid pulse rate may be a sign of severe blood loss or of disease within the heart itself. Irregularity of the pulse (arrhythmia) is an important indicator of heart malfunction.

The respiratory rate (rate of breathing) is modified by disease. Persons with fever have an increased respiratory rate (hyperventilation), which serves to lower body temperature (this rapid breathing is analogous to the panting of a dog). Hyperventilation is a common response to painful stress. Any condition leading to acidosis (lowering of body pH) similarly drives the respiratory rate upward. Diseases of the lungs—with the accompanying inability to oxygenate the blood adequately—have a similar effect.

Temperature, pulse, and respiratory rate—called the vital signs—may be important manifestations of disease. A fourth vital sign, blood pressure, is equally significant. Among other things, it indicates the amount of blood in circulation. A decrease in circulating blood volume, as is seen with severe bleeding, lowers the blood pressure and deprives the tissues of adequate blood flow. Reflexes are initiated that compensate for the reduced blood volume and blood pressure. The heart rate increases and compensates to some extent for the sudden reduction in blood volume and pressure; at the same time, peripheral blood vessels in such areas as the abdomen constrict, tending to divert the reduced blood volume to the more vital areas such as the brain and head. Unusual elevation of pressure (hypertension) is a disease by itself.

Causes of edema

Fluid and electrolyte imbalances may be further consequences of homeostatic failure and additional significant manifestations of disease. The causes of these abnormalities are complex. Edema, or swelling, results from shifts in fluid distribution within body tissues. Edema may be localized, as when the leg veins are narrowed or obstructed by some disease process. The pressure of the blood in the distended veins rises, and fluid is driven out of the vessels into the tissues, causing swelling of the extremity. Generalized edema is seen in renal (kidney) disease that causes abnormal retention of sodium and water. Heart failure is an additional cause of generalized edema, usually most manifest as swollen feet and ankles. Alterations such as dehydration, hyperventilation, and tissue destruction can all lead to varying fluid and electrolyte derangements. The levels of the serum electrolytes (sodium, potassium, bicarbonate, chloride), determined relatively easily in the laboratory, provide the physician with valuable clues to deranged homeostasis induced by disease.

Finally, the determination of body pH and a number of blood tests designed to evaluate adequate (or inadequate) metabolic regulation provide diagnostic clues of homeostatic failure. These tests include determination of the levels of the blood glucose, blood urea nitrogen, and serum protein.

The disease diabetes mellitus provides an excellent example of failure of the homeostatic mechanisms. Diabetes is a common disease of metabolic-endocrine (ductless gland) origin involving a relative or absolute deficiency of insulin, a hormone that plays a major role in carbohydrate metabolism. Any or all of the homeostatic derangements can be found in this disease. Patients with a severe form of diabetes may at one time be dehydrated because of obligatory excretion of water (osmotic diuresis), be acidotic because of formation of increased amounts of keto acids derived from the oxidation of free fatty acids, be hyperventilating as a result of the acidosis, be comatose because of high levels of blood glucose, have a weak pulse because of severe dehydration, have electrolyte abnormalities, and so on. The signs and symptoms are numerous, all illustrating the interdependence of the homeostatic mechanisms, which, when not functioning properly, provide the manifestations of disease.

At the most elemental level, disease develops when any disruptive or adverse influence overcomes the homeostatic and defensive controls of the body. As will be seen, there are numerous influences that can tip the scales of health toward disease. Viruses and bacteria are obvious threats to health. There are a great many others, some so subtle as to be poorly understood. The following section focuses on the causes of disease rather than on a detailed description of each entity. It represents one method of classification. There is considerable overlap in categories; certain diseases grouped as metabolic-endocrine in origin could also be classified as diseases of genetic origin. Indeed, the interdependence of the organ systems, the metabolic pathways, and the defense systems renders finite classification in medicine difficult. The human body acts as a unit—an individual—both in health and in disease.

Overlap in causal categories

The causes of disease

The search for the causes (etiologies) of human diseases goes back to antiquity. Hippocrates, a Greek physician of the 4th and 5th centuries BC, is credited with being the first to adopt the concept that disease is not a visitation of the gods but rather is caused by earthly influences. Scientists have since continually searched for the causes of disease and, indeed, have discovered the causes of many.

In the development of a disease (pathogenesis) more is involved than merely exposure to a causative agent. A room full of people may be exposed to a sufferer from a common cold, but only one or two may later develop a cold. Many host factors determine whether the agent will induce disease or not. Thus, in the pathogenesis of disease, the resistance, immunity, age, and nutritional state of the person exposed, as well as virulence or toxicity of the agent and the level of exposure, all play a role in determining whether disease develops.

In the following sections the many types of human disease will be divided into categories, and in each only a few examples will be given to establish the nature of the process. These categories are divided on the basis of the presumed etiology of the disease. Many diseases are still of unknown (idiopathic) origin. With others the cause may be suspected but not yet definitively proved.

DISEASES OF GENETIC ORIGIN

Certain human diseases result from mutations in the genetic complement (genome) contained in the deoxyribonucleic acid (DNA) of chromosomes. A gene is a discrete linear sequence of nucleotide bases (molecular units) of the DNA that codes for, or directs, the synthesis of a protein, and there may be as many as 100,000 genes in the human

genome. Proteins, many of which are enzymes, carry out all cellular functions. Any alteration of the DNA may result in the defective synthesis and subsequent malfunctioning of one or more proteins. If the mutated protein is a key enzyme in normal metabolism, the error may have serious or fatal consequences. More than 5,000 distinct diseases have been ascribed to mutations that result in deficiencies of critical enzymes.

Mutations are classified on the basis of the extent of the alteration. Large mutations, which include alterations to chromosome structure and number, are relatively rare because most cause such major disruptions to development that the fetus is naturally aborted. However, certain alterations are not so immediately lethal, and the fetus can survive with a characteristic disorder. Down syndrome is one such case. It involves an error in the division of chromosome 21 that results in trisomy (three copies of a chromosome instead of two are inherited), bringing the total number of chromosomes to 47 instead of 46. Many characteristics such as distinctive facial features and mental retardation result from the presence of this extra chromosome. Smaller mutations are more common and include point mutations, in which substitution of a single nucleotide base occurs, and deletion or insertion mutations, which involve several bases. Point, deletion, and insertion mutations may cause an abnormal protein to be synthesized or may prevent the protein from being made at all.

Mutations that occur in the DNA of somatic (body) cells cannot be inherited, but they can cause congenital malformations and cancers (see below *Abnormal growth of cells*); however, mutations that occur in germ cells—*i.e.,* the gametes, ova and sperm—are transmitted to offspring and are responsible for inherited diseases. Each gamete contributes one set of chromosomes and therefore one copy (allele) of each gene to the resultant offspring. If a gene bearing a mutation is passed on, it may cause a genetic disorder.

Genetic diseases caused by a mutation in one gene are inherited in either dominant or recessive fashion. In dominantly inherited conditions, only one mutant allele, which codes for a defective protein or does not produce a protein at all, is necessary for the disorder to occur. In recessively inherited disorders, two copies of a mutant gene are necessary for the disorder to manifest; if only one copy is inherited, the offspring is not affected, but the trait may continue to be passed on to future offspring. In addition to dominant or recessive transmission, genetic disorders may be inherited in an autosomal or X-linked manner. Autosomal genes are those not located on the sex chromosomes, X and Y; X-linked genes are those located on the X chromosomes that have no complementary genes on the Y chromosome. Females have two copies of the X chromosome, but males have an X and a Y chromosome. Because males have only one copy of the X chromosome, any mutation occurring in a gene on this chromosome will be expressed in male offspring regardless of whether its behaviour is recessive or dominant in females. Autosomal dominant disorders include Huntington's chorea, a degenerative disease of the nervous system that usually does not develop until the carrier is between 30 and 40 years of age. The delayed onset of Huntington's chorea allows this lethal gene to be passed on to offspring. Autosomal recessive diseases are more common and include cystic fibrosis, Tay-Sachs disease, and sickle cell anemia. X-linked dominant disorders are rare, but X-linked recessive diseases are relatively common and include Duchenne's muscular dystrophy and hemophilia A.

Congenital genetic disorders

Most genetic disorders can be detected at birth because the child is born with characteristic defects. Thus these abnormalities are congenital (existing at birth) genetic disorders. A few genetic defects, such as Huntington's chorea mentioned above, do not become manifest until later in life. Hence it may be said that most but not all genetic diseases are congenital.

Conversely, some congenital diseases are not genetic in origin; instead they may arise from some direct injury to the developing fetus. If a woman contracts the viral disease German measles (rubella) during pregnancy, the virus may infect the fetus and alter its normal development, leading to some malformations, principally of the heart. These malformations constitute a congenital disease that is not genetic.

Further confusion often arises over the terms genetic and familial. A familial disease is hereditary, passed on from one generation to the next. It resides in a genetic mutation that is transmitted by mother or father (or both) through the gametes to their offspring. Not all genetic disorders are familial, however, because the mutation may arise for the first time during the formation of the gametes or during the early development of the fetus. Such an infant will have some genetic abnormality, though the parents themselves do not. Down syndrome is an example of a genetic disease that is not familial.

Factors relating to genetic injury. The causes of mutations are still poorly understood. Certain factors, however, are thought to be important. Maternal age plays an important role in predisposing toward genetic injury. The frequency of Down syndrome and of congenital malformations increases with the age of the mother. This may be so for a variety of reasons. Unlike men, who produce new sperm continually, women are born with all the eggs (ova) they will ever have. Thus the eggs are exposed to the same internal and external agents that the woman comes in contact with. The longer the exposure to such factors (*i.e.,* the older the mother), the greater the chance of genetic injury to the ova. A paternal contribution to the disease also has been discovered—roughly 25 percent of cases may be caused by extra chromosomal material from the father. At present, the nature of the factors responsible for impaired division of chromosomes remains unknown.

Congenital defects and age of mother

Radiation is a well-recognized cause of chromosomal damage. The survivors of the atomic bomb blasts in Japan in 1945 have shown definite chromosomal abnormalities in certain types of their circulating white blood cells. Indeed, a higher incidence of leukemia (a form of cancer of white cells), as well as other cancers, has been reported in this population, suggesting that the chromosomal changes may have played some role in the induction of the disease (see also RADIATION: *Biologic effects of ionizing radiation*).

Viruses have been shown to cause mutations in human cells when the cells are grown in tissue culture, but there is no clear evidence that viral infections can cause genetic injury in humans. Instead, current evidence suggests that the oncogenic viruses implicated in some human cancers facilitate genetic mutations rather than cause them directly.

The induction of DNA mutations in cells by drugs and chemicals is complex. It involves metabolism of the drug by detoxification enzymes into reactive intermediates that damage DNA. The mutations that remain are those not removed by DNA repair enzymes. In contrast to viruses, drugs and chemicals have been shown to cause mutations not only in human cells in culture but also in a living host.

Heredity and environment. Diseases can be spread across a wide spectrum, with predominantly genetic diseases at one extreme of the spectrum and diseases of largely environmental origin at the other. In the genetic part of the spectrum are diseases such as Turner's syndrome; in the environmental part are infectious diseases and chemical poisoning. Between these two extremes lie most human diseases—those with both genetic and environmental causative influences that are significant. Indeed, even at the very extreme ends of the spectrum both factors play some role. The genetic constitution dictates in part the host's response to environmental challenges. Similarly, environmental factors play significant roles in the manifestation of genetically induced disease. Sickle cell anemia, for example, an inherited disease characterized by abnormal red blood cells and hemoglobin, is seriously exacerbated by low levels of oxygen in the air.

Familial tendencies to disease

Furthermore, there are many disorders in which there is a familial tendency to develop the disease but no formal pattern of inheritance has been delineated. Many forms of cancer, high blood pressure, arthritis, and obesity, for example, seem to have a familial tendency. Although the exact roles of environmental and genetic factors are unknown in all these diseases, it is strongly felt that both factors contribute to the disease process.

CHEMICAL AND PHYSICAL INJURY

Chemical injury: poisoning. A poison is any substance that can cause illness or death when ingested in small quantities. This definition excludes the multitude of substances that cause damage if ingested in large quantities. For example, even oxygen and glucose, so crucial to life, are toxic to cells when administered at high concentrations.

There are several considerations to keep in mind when one discusses poisoning. The first of these, as already suggested, is the degree of toxicity. A substance with a very high toxicity (such as cyanide) need be taken only in minute amounts to cause serious harm or death.

A second consideration is the mechanism by which a poison operates. Each poison acts at particular sites in the cell that are critical for the maintenance of homeostasis. These sites include the genome, whose expression dictates cell structure and function, and the cell membrane, which regulates ion transport, energy metabolism, and synthesis of vital proteins. Each poison also has a characteristic ability to cause damage at particular sites within the body, such as the liver, kidneys, or central nervous system.

A third factor is the body's ability to eliminate the substance. Some chemicals, rapidly excreted in the urine, must act quickly while they remain transiently in the body. Others are poorly eliminated, and, because of this, a chronic ingestion of nontoxic amounts leads to a buildup in the body that can reach toxic levels. Lead poisoning is a good example of this phenomenon.

The route of entry is also important. Many substances are harmless when eaten but become deadly if injected into a vein. There are chemicals and drugs that are highly reactive and interact directly with an important cellular component to cause cell injury or death. Other chemicals or drugs that are not toxic per se become so following their metabolic conversion to toxic intermediates by the host. Similarly, the chemical form of a substance affects its action on the body. Metallic mercury, as found in thermometers, is harmlessly excreted, whereas the chloride salt of the same substance is deadly.

Finally, the condition of the host, the recipient of the poison, is an important consideration. A dose of aspirin (acetylsalicylic acid) that is harmless to an adult may be poisonous to an infant. Similarly, an elderly person's tolerance of a substance may be much lower than that of a healthy young adult.

A wide variety of poisons exist, among which a few stand out as being the most commonly encountered in medical practice. Some are of relatively low toxicity but are important because of their widespread use. Many physicians consider aspirin the most dangerous poison because of its commonplace use and abuse and because it is the leading cause of poisoning in children. In the following paragraphs three groups of agents will be presented: (1) organic chemicals, (2) inorganic chemicals, and (3) drugs.

Organic chemicals. Among the organic chemicals commonly encountered in instances of poisoning are two forms of alcohol, ethyl alcohol (ethanol) and methyl alcohol (methanol). Ethyl alcohol is the form found in most alcoholic beverages. Methyl alcohol, or wood alcohol, is used for a variety of household purposes.

Ethyl alcohol poisoning

Acute ethyl alcohol poisoning is encountered after ingestion of large quantities over a relatively short time. The alcohol is quickly absorbed from the gastrointestinal tract, and high blood levels can be achieved in a remarkably short time. Ethyl alcohol acts principally as a central-nervous-system depressant and, fortunately, stupor usually results before fatal doses can be reached. The difference in blood levels between intoxication and fatal stupor is very slight, however, and death may result with the ingestion of large quantities of alcohol from depression of the respiratory centre in the brain.

Carbon monoxide is a nonirritating, inert gas without colour, taste, or odour. A poison responsible for a large number of accidental and suicidal deaths, it is one of the chemical products of any combustion of organic material. Inhalation of a 1 percent concentration can be fatal within 10 to 20 minutes. Carbon monoxide acts as an internal asphyxiant causing oxygen starvation of tissues. It should be noted that exposure to even low concentrations can result in the slow accumulation of this poison over hours, days, or weeks, leading very gradually to toxic or fatal levels.

Inorganic chemicals. The inorganic chemicals most commonly responsible for poisonings in the United States are cyanide, mercury, arsenic, and lead. While the last three often appear in chemical forms that are quite harmless, it is the soluble salts of the substances that are poisons.

Cyanide poisoning

Cyanide is a dangerous substance in any form. It may occur in the form of hydrocyanic gas or as solid compounds such as potassium cyanide. It is one of the most lethal poisons known; an amount of 0.2 gram (0.007 ounce) administered to a 70-kilogram (154-pound) human causes death within minutes. Like carbon monoxide, it acts as a cellular asphyxiant.

Mercury in the pure metallic form is rather harmless, but the salt of the same substance, notably mercuric chloride, is a deadly poison. As little as 0.1 gram is enough to cause damage to body tissues, and 2 grams can cause death in a 70-kilogram person. This agent causes extensive tissue damage wherever high concentrations of the poison are encountered. When the substance is swallowed, the stomach represents the portal of entry. The mercuric chloride is partially absorbed into the blood, and this portion is excreted through the urine. The remainder affects organs in the digestive tract, principally the stomach and the colon, and the kidneys. Mercuric salts cause death of cells by precipitating the proteins within the cells, a form of cell injury called coagulative necrosis. With careful treatment, affected persons survive with full recovery. Chronic ingestion of smaller amounts of mercuric salts, as is seen in some industrial settings, can result in disease involving the mouth, skin, and nervous system.

Arsenic is contained in many items used around the house. Both odourless and tasteless compounds of arsenic are found in some rat poisons, plant sprays, paints, and other household preparations. Many of these household staples are ingested accidentally by children. Principally affected by arsenic are the blood vessels and the central nervous system; vascular collapse and depression of the central nervous system can be followed by coma and death within hours after ingestion.

Lead salts

The soluble salts of inorganic lead are also strong systemic poisons. They may accumulate within the body over a long period until toxic levels are reached and cell damage ensues. These salts were at one time commonly found in paints, and lead poisoning was frequently seen in children who chewed on their painted cribs or woodwork. Legislation in many countries has outlawed the use of lead-base paints for infants' furniture. Other forms of poisoning are incurred through industrial exposure and ingestion of water from lead pipes. Lead poisoning damages red blood cells and leads to hemolysis (rupturing of red blood cells) with resulting anemia. In the brain, lead accumulation causes the degeneration of nerve cells. This produces such manifestations as mental depression, psychoses, convulsions, and even coma and death. If an early fatality does not occur, the lead is slowly excreted and complete recovery may be anticipated.

Drugs. Drugs are another important cause of poisoning. It is a pharmacological principle that, for any therapeutic gain derived from a drug, a price is paid. There are few drugs used today that have no side effects (*i.e.,* effects unintended when the drug is administered). Although these side effects may be harmless and inconsequential, certain drugs have side effects that are potent. Similarly, a drug may be useful in a certain dose range but harmful when larger doses are taken. Morphine, for example, is an excellent drug for the control of severe pain, but it can depress respiration, and too much of it can cause death. All drugs are, therefore, potentially harmful.

Barbiturates and salicylates are the major drugs commonly found to cause serious illness from overingestion. Barbiturates affect the central nervous system almost exclusively. With toxic levels, the vital centres located within the midbrain are depressed; this leads to profound coma, depression of respiration, oxygen starvation of the tissues, and even shock. The identification of barbiturate poisoning relies almost exclusively on finding the substance in the blood or urine, because there is little anatomic change

in tissues. Treatment is directed toward getting the drug out of the system as quickly as possible, either by inducing copious urinary excretion of the drug or by the use of the artificial kidney—a process called hemodialysis.

Aspirin

Aspirin, or acetylsalicylic acid, is a drug that deserves special mention because it is such a common household item and often within the reach of small children. Approximately 10 to 30 grams of aspirin can be fatal in adults, and much smaller amounts can be fatal in children. (A single aspirin tablet of standard size contains approximately one-third gram.) There are many signs and symptoms associated with salicylate poisoning, including headaches, drowsiness, dyspepsia, nausea, vomiting, sweating, and thirst. Salicylate poisoning is an acute medical emergency. Rigorous medical treatment is demanded, and use of the artificial kidney is often required.

Physical injury. Physical injuries include those caused by mechanical trauma, heat and cold, electrical discharges, changes in pressure, and radiation. Mechanical trauma is an injury to any portion of the body from a blow, crush, cut, or penetrating wound. The complications of mechanical trauma are usually related to fracture, hemorrhage, and infection. They do not necessarily have to appear immediately after occurrence of the injury. Slow internal bleeding may remain masked for days and lead to an eventual emergency. Similarly, wound infection and even systemic infection are rarely detectable until many days after the damage. All significant mechanical injuries must therefore be kept under observation for days or even weeks.

Injuries from cold or heat. Among physical injuries are injuries caused by cold or heat. Prolonged exposure of tissue to freezing temperatures causes tissue damage known as frostbite. Several factors predispose to frostbite, such as malnutrition leading to a loss of the fatty layer under the skin, lack of adequate clothing, and any type of insufficiency of the peripheral blood vessels, all of which increase the loss of body heat.

When the entire body is exposed to low temperatures over a long period, the result can be alarming. At first blood is diverted from the skin to deeper areas of the body, resulting in anoxia (lack of oxygen) and damage to the skin and the tissues under the skin, including the walls of the small vessels. This damage to the small blood vessels leads to swelling of the tissues beneath the skin as fluid seeps out of the vessels. When the exposure is prolonged, it leads eventually to cooling of the blood itself. Once this has occurred, the results are catastrophic. All the vital organs become affected, and death usually ensues.

Three categories of burns

Burns may be divided into three categories depending on severity. A first-degree burn is the least destructive and affects the most superficial layer of skin, the epidermis. Sunburn is an example of a first-degree burn. The symptoms are pain and some swelling. A second-degree burn is a deeper and hence more severe injury. It is characterized by blistering and often considerable edema (swelling). A third-degree burn is extremely serious; the entire thickness of the skin is destroyed, along with deeper structures such as muscles. Because the nerve endings are destroyed in such burns, the wound is surprisingly painless in the areas of worst involvement.

The outlook in burn injuries is dependent on the age of the victim and the percent of total body area affected. Loss of fluid and electrolytes and infection associated with loss of skin provide the major causes of burn mortality.

Electrical injuries. The injurious effects of an electrical current passing through the body are determined by its voltage, its amperage, and the resistance of the tissues in the pathway of the current. It must be emphasized that exposure to electricity can be harmful only if there is a contact point of entry and a discharge point through which the current leaves the body. If the body is well insulated against such passage, at the point of either entry or discharge, no current flows and no injury results. The voltage of current refers to its electromotive force, the amperage to its intensity. With high-voltage discharges, such as are encountered when an individual is struck by lightning, the major effect is to disrupt nervous impulses; death is usually caused by interruption of the regulatory impulses of the heart. In low-voltage currents, such as are more likely to be encountered in accidental exposure to house or industrial currents, death is more often due to the stimulation of nerve pathways that cause sustained contractions of muscles and may in this way block respiration. If the electrical shock does not produce immediate death, serious illness may result from the damage incurred by organs in the pathway of the electrical current passing through the body.

Pressure-change injuries. Physical injuries from pressure change are of two general types: (1) blast injury and (2) the effects of too-rapid changes in the atmospheric pressure in the environment. Blast injuries may be transmitted through air or water; their effect depends on the area of the body exposed to the blast. If it is an air blast, the entire body is subject to the strong wave of compression, which is followed immediately by a wave of lowered pressure. In effect the body is first violently squeezed and then suddenly overexpanded as the pressure waves move beyond the body. The chest or abdomen may suffer injuries from the compression, but it is the negative pressure following the wave that induces most of the damage, since overexpansion leads to rupture of the lungs and of other internal organs, particularly the intestines. If the blast injury is transmitted through water, the victim is usually floating, and only that part of the body underwater is exposed. An individual floating on the surface of the water may simply be popped out of the water like a cork and totally escape injury.

Blast injuries

Decompression sickness is a disease caused by a too-rapid reduction in atmospheric pressure. Underwater divers, pilots of unpressurized aircraft, and persons who work underwater or below the surface of the Earth are subject to this disorder. As the atmospheric pressure lessens, dissolved gases in the tissues come out of solution. If this occurs slowly, the gases diffuse into the bloodstream and are eventually expelled from the body; if this occurs too quickly, bubbles will form in the tissues and blood. The oxygen in these bubbles is rapidly dissolved, but the nitrogen, which is a significant component of air, is less soluble and persists as bubbles of gas that block small blood vessels. Affected individuals suffer excruciating pain, principally in the muscles, which causes them to bend over in agony—hence the term "bends" used to describe this disorder.

Radiation injury. Radiation can result in both beneficial and dangerous biological effects. There are basically two forms of radiation: particulate, composed of very fast-moving particles (alpha and beta particles, neutrons, and deuterons), and electromagnetic radiation such as gamma rays and X rays. From a biological point of view, the most important attribute of radiant energy is its ability to cause ionization—to form positively or negatively charged particles in the body tissues that it encounters, thereby altering and, in some cases, damaging the chemical composition of the cells. DNA is highly susceptible to ionizing radiation. Cells and tissues may therefore die because of damage to enzymes, because of the inability of the cell to survive with a defective complement of DNA, or because cells are unable to divide. The cell is most susceptible to irradiation during the process of division. The severity of radiation injury is dependent on the penetrability of the radiation, the area of the body exposed to radiation, and the duration of exposure, variables that determine the total amount of radiant energy absorbed.

When the radiation exposure is confined to a part of the body and is delivered in divided doses, a frequent practice in the treatment of cancer, its effect depends on the vulnerability of the cell types in the body to this form of energy. Some cells, such as those that divide actively, are particularly sensitive to radiation. In this category are the cells of the bone marrow, spleen, lymph nodes, sex glands, and lining of the stomach and intestines. In contrast, permanently nondividing cells of the body such as nerve and muscle cells are resistant to radiation. The goal of radiation therapy of tumours is to deliver a dosage to the tumours that is sufficient to destroy the cancer cells without too severely injuring the normal cells in the pathway of the radiation. Obviously, when an internal

cancer is treated, the skin, underlying fat, muscles, and nearby organs are unavoidably exposed to the radiation. The possibility of delivering effective doses of radiation to the unwanted cancer depends on the ability of the normal cells to withstand the radiation. However, as is the case in drug therapy, radiation treatment is a two-edged sword with both positive and negative aspects.

Finally, there are probable deleterious effects of radiation in producing congenital malformations, certain leukemias, and possibly some genetic disorders (see RADIATION: *Biologic effects of ionizing radiation*).

DISEASES OF IMMUNE ORIGIN

The immune system protects against infectious disease, but it may also at times cause disease. Disorders of the immune system fall into two broad categories: (1) those that arise when some aspect of the host's immune mechanism fails to prevent infection (immune deficiencies) and (2) those that occur when the immune response is directed at an inappropriate antigen, such as a noninfectious agent in an allergic reaction, the body's own antigens in an autoimmune response, or the cells of a transplanted organ in graft rejection.

Types of immune disorders

Immune deficiencies. The immune system may fail to function for many reasons. Many immunodeficiency disorders are caused by a genetic defect in some component of the system and thus usually manifest early in life. Some deficiencies, however, are acquired through the action of infectious agents such as viruses, through the action of immunosuppressive agents used to treat various medical conditions, and through the effects of certain disease processes such as cancer. Both inherited and acquired immune deficiencies suppress one or many aspects of the immune response, rendering the affected individual unable to resist infection unless treated by administration of immunoglobulins or by bone marrow transplant.

Inherited immune disorders undermine the immune response in a variety of ways: B lymphocytes may be unable to produce antibodies, phagocytes may be unable to digest microbes, or specific complement components may not be produced. Severe combined immunodeficiency (SCID), a condition that arises from several different genetic defects, disrupts the functioning of both the humoral and cell-mediated immune responses.

Acquired immune deficiency syndrome (AIDS) is caused by infection with the human immunodeficiency virus (HIV), which destroys a certain type of T lymphocyte, the helper T cell. An infected individual is susceptible to a variety of infectious organisms, including those called opportunistic pathogens, which may live benignly in the human body and cause disease only when the immune system is suppressed. Certain diseases such as Kaposi's sarcoma and *Pneumocystis carinii* pneumonia, which until recently were rarely encountered by clinicians, have become prevalent in the AIDS population and are often the cause of mortality.

Immune responses in the absence of infection. *Allergies.* The immune system may react to any foreign substance, and consequently it can respond to innocuous materials in the same way that it responds to infectious agents. If the foreign material poses no threat to the individual, an immune response is unnecessary, but it nevertheless may ensue. This misplaced response is called an allergy, or hypersensitivity, and the foreign material is referred to as an allergen. Common allergens include pollen, dust, bee venom, and various foods such as shellfish. What causes one person and not another to develop an allergy is not completely understood.

The allergic response

An allergic response occurs in the following manner. On first exposure to the allergen, the person becomes sensitized to it—that is, develops antibodies and specific T cells to the allergen. An allergic reaction does not usually accompany this initial event. When reexposure occurs, however, symptoms of the allergic response appear. These symptoms range from the mild response of sneezing and a runny nose to the sometimes life-threatening reaction of anaphylaxis, or anaphylactic shock, symptoms of which include vascular collapse and potentially fatal respiratory distress.

Allergic reactions exhibit different symptoms depending on which immune mechanisms are responsible. On the basis of this criterion, they can be categorized into four types, the first three of which involve antibodies and occur in a matter of minutes or hours. Type IV hypersensitivity, unlike the other reactions, does not involve antibodies but instead is mediated by T cells. In these reactions, also called delayed-type because they arise in a matter of days rather than minutes or hours, T cells either activate a local inflammatory reaction, which can cause extensive tissue damage, or they kill tissue cells directly. Chronic inflammation characteristic of many autoimmune disorders, such as chronic thyroiditis, results from this reaction. With the exception of the type I response, all responses are seen in both allergies and autoimmune disorders.

Autoimmune disorders. Immune responses can be mounted against proteins that belong to the host, giving rise to autoimmune diseases. Although the immune system naturally generates antibodies to its own cells, mechanisms exist to keep this activity in check. Two mechanisms that prevent the immune system from mounting an attack against the host's own tissues have been identified. The first involves the elimination of self-reactive lymphocytes during their development and maturation in the thymus, a lymphoid organ in the chest. Self-reactive lymphocytes present in these cell populations are destroyed when they encounter the self-antigen to which they react. Because this protective selection process is not highly efficient, some self-reactive lymphocytes survive, exit the thymus, and enter the blood and tissues. Outside the thymus a second line of defense against immune self-destruction is afforded in which self-reactive lymphocytes lose their ability to react to self-antigens when they are encountered in blood and tissues. This state is referred to as immunologic ignorance. Autoimmune diseases arise when this mechanism fails and self-reactive lymphocytes are activated by self-antigens in the host's own tissues, often with devastating effects. Systemic lupus erythematosus, thyroiditis, insulin-dependent diabetes mellitus, and rheumatoid arthritis are examples of this type of disorder.

Graft rejection. Transplantation of organs and cells from one individual to another has become an important medical treatment. As are other forms of therapy, it is accompanied by certain risks. Each individual's cells have a spectrum of genetically determined cell surface protein antigens, the major histocompatibility complex (MHC) antigens, or human leukocyte antigens as they are referred to in humans. MHC antigens determine a person's tissue type just as red blood cell antigens determine blood type. There are two classes of MHC antigens: class I molecules, encoded by three genes, and class II molecules, encoded by four possible sets of genes. Each of these genes has many alternative forms, and thus the probability of any two individuals—aside from siblings, especially identical twins—having the same form of each gene is extremely small. Even parents will have different tissue antigens from their children.

These differences in tissue antigens pose an obstacle to transplantation because it is highly likely that foreign donor tissue will introduce antigens in the recipient that will trigger an immune response leading to tissue death and rejection. However, by careful matching of the MHC type of donor and recipient, rejection can be diminished or avoided. Because perfect matching is possible only between identical twins or very close relatives, many transplants occur between less closely matched tissue types, and success is achieved with the administration of powerful immunosuppressive drugs.

DISEASES OF BIOTIC ORIGIN

Factors in infection. *Infectious agents.* Biotic agents include life-forms that range in size from the smallest virus, measuring approximately 20 nanometres (0.000 000 8 inch) in diameter, to tapeworms that achieve lengths of 10 metres (33 feet). These agents are commonly grouped as viruses, rickettsiae, bacteria, fungi, and parasites. The disease that these organisms cause is only incidental to their struggle for survival. Most of these agents do not require a human host for their life cycles. Many survive readily

in soil, water, or lower animal species and are harmless to humans. Other living organisms, which require the temperature range of endothermic (warm-blooded) animals, may flourish on the skin or in the secretions of fluids of the mouth or intestinal tract but do not invade tissue or cause disease under normal conditions. Thus there is a distinction to be made between infection and disease.

All animals are infected with biotic agents. Those agents that do not cause disease are termed nonpathogenic, or commensal. Those that invade and cause disease are termed pathogenic. *Streptococcus viridans* bacteria, for example, are found in the throats of more than 90 percent of healthy persons. In this area they are not considered pathogenic. The same organism cultured from the bloodstream, however, is highly pathogenic and usually indicates the presence of the disease subacute bacterial endocarditis (chronic bacterial invasion of the valves of the heart). In order for such nonpathogenic agents to achieve pathogenicity, they obviously must overcome the defenses of the host. Most biotic agents require a portal of entry through the intact skin or mucosal linings of the body. They must be present in sufficient number to escape the phagocytes. They must be capable of surviving the inflammatory and immune response. Ultimately, to induce disease, they must have sufficient virulence and invasiveness to cause significant tissue injury.

Invasiveness

Invasiveness and virulence. Invasiveness is the capability of penetrating and spreading throughout tissues. Remarkably, little is known of the factors that condition it. In a few instances enzymes produced by biotic agents have been identified that are capable of breaking down the integrity of the supporting tissues of the body, thereby preparing a pathway for the spread of the organism.

Only very few bacteria release such enzymes, however, and there are marked differences in invasiveness to be found among the various types of bacteria. The organism that causes diphtheria (*Corynebacterium diphtheriae*), for example, is capable of invading only the surface cells of the mouth and throat. The disease that results is caused by the production of a powerful exotoxin (a chemical substance produced by the organism and released into the surrounding tissues) that is absorbed into the bloodstream from the local infection within the throat. This exotoxin causes major damage in the heart and the nervous system. The diphtheria bacillus, therefore, is an example of a serious infection in which the organism has low invasiveness. In contrast, the bacterium that causes syphilis (*Treponema pallidum*) has a high degree of invasiveness. It is one of the rare biotic agents that are capable of penetrating intact skin and mucosal linings of the body.

The invasiveness of viruses undoubtedly is facilitated by their extremely small size, but, because of this size, the exact mechanism is difficult to study. In the case of fungi and parasites, the invasiveness is related to the life cycle of the organism. The formation of tiny spores by fungi and the smaller reproductive forms of the parasites provide vehicles by which infection may be drawn into the lungs or may pass through tiny defects in the skin or mucosal linings of the various openings and tracts of the body.

In general, virulence is the degree of toxicity or the injury-producing potential of a microorganism. The words virulence and pathogenicity are often used interchangeably. The virulence of bacteria usually relates to their capability of producing a powerful exotoxin or endotoxin. Invasiveness also adds to an organism's virulence by permitting it to spread.

Predisposition of the host. Up to this point, diseases caused by biotic agents have been considered in terms of the role of the invader. Equally important is the role of the host, the individual who contracts the disease. Any infectious disease is a test between the invader and the defender. Virulent organisms may be capable of inducing serious illness even in the most robust. The converse is perhaps more important. The weak host is prey to many forms of biotic infection, even those of low virulence and invasiveness. Some of the more important of the many factors that condition the level of resistance to biotic infection in the individual are age, with infancy and old age being times of maximum vulnerability; poor nutrition; genetic disorders and immunosuppressive agents, such as the human immunodeficiency virus, that compromise the immunologic system; and metabolic disorders such as diabetes that increase vulnerability to infectious agents.

Drugs as factors predisposing to disease

Therapeutic agents, paradoxically, also have become important factors in predisposing to disease of biotic origin and indeed in altering the incidence patterns of infectious disease. The drugs that are principally involved include those used to suppress the immune response, as well as the host of antimicrobial and antibiotic agents now employed in the treatment of infectious disease. Immunosuppressive drugs are used to block the immune response in patients about to receive an organ transplant and in the treatment of the autoimmune diseases, but such treatment renders the patient vulnerable to attack by biotic agents. Indeed, these immunologically compromised persons become susceptible to organisms of extremely low virulence.

Antimicrobial drugs also have drawbacks as well as benefits. A patient suffering from a streptococcal disease, for example, may appropriately be treated with penicillin. Certain strains of staphylococci, however, are resistant to penicillin. Although the streptococcal organisms, as well as other commensals, may be eradicated by the antibiotic, the resistant staphylococci begin to proliferate, possibly because the competition with other bacteria for nutrients and food supply has been removed. In this noncompetitive situation they may cause disease. More powerful antibiotics may destroy all bacteria, including staphylococci, but permit the unrestrained proliferation of fungi and other agents of low virulence that are nonetheless resistant to the antibiotic. Thus antibiotics have changed the entire frequency pattern of biotic disease. Organisms that have proved to be more resistant to antibiotics have become the more common causes of serious clinical infection. For this reason certain forms of drug-resistant bacteria that include *Escherichia coli, Aerobacter aerogenes, Pseudomonas aeruginosa,* and strains of *Proteus* as well as fungi have emerged as the important biotic causes of death.

Types of viral disease

Viral diseases. Of the many existing viruses, a few are of great importance as causes of human sickness. They are responsible for such diseases as smallpox, poliomyelitis, encephalitis, influenza, yellow fever, measles, and mumps and such minor disorders as warts and the common cold.

Viruses may survive for some time in the soil, in water, or in milk, but they cannot multiply unless they invade or parasitize living cells. Certain viruses proliferate within the host cells and accumulate in sufficient number to cause rupture and death of the cells. Others multiply within the cell body and compete with the host for nutrition or vital constituents of the cell's metabolism. Both types of viruses are said to be cytotoxic.

Viral agents, particularly those capable of producing tumours in humans and lower animals, flourish within cells and stimulate the cells to active growth. These viruses are referred to as oncogenic (tumour-producing). The number of oncogenic viruses that cause tumours in lower animals is large. In humans, several DNA viruses and one RNA virus have been implicated strongly in the induction of a variety of tumours (see CANCER).

Most viral infections occur in childhood. This age distribution has been explained on immunologic grounds. Viruses usually induce a firm and enduring immunity. On first exposure to a virus, children may or may not contract the disease, depending on their resistance, the size of the infective dose of virus, and many other variables. Those who contract the disease, as well as those who resist the infection, develop a permanent immunity to any further exposure. By either pathway, as children grow older they progressively gather protection against viral infections. Consequently, the incidence of these infections falls in adulthood and later life. The frequency of common colds is explained on the grounds that a host of different viral agents all induce similar respiratory infections, and, while a single attack confers immunity against the specific causative agent, it provides no protection against the rest.

Viral diseases are resistant to antibiotics and other antimicrobial agents. This point is made because of a distressing tendency among individuals to take penicillin or another antibiotic for a common cold.

Micro-organisms as cause

Rickettsial diseases. Human rickettsial diseases are caused by microorganisms that fall between viruses and bacteria in size. These minute agents are barely visible under the ordinary light microscope. Like viruses, they multiply only within the cells of susceptible hosts. They are found in nature in a variety of ticks and lice and, when transmitted to humans by the bite of one of these arthropods, usually cause acute febrile (fever-producing) illnesses, most of which are characterized by skin rashes. Rocky Mountain spotted fever, a systemic rickettsial infection, invades and kills the cells lining blood vessels and causes hemorrhage, inflammation, blood clots, and extensive tissue death; if untreated, it is fatal in about 20 to 30 percent of cases.

Bacterial diseases. The diseases produced by bacteria are the most common of infectious biotic diseases. They range from trivial skin infections to such devastating disorders as bubonic plague and tuberculosis. Various types of pneumonia; infections of the cerebrospinal fluid (meningitis), the liver, and the kidneys; and the sexually transmitted diseases syphilis and gonorrhea are all forms of bacterial infection.

All bacteria induce disease by one of three methods: (1) the production of an exotoxin, a harmful chemical substance that is secreted or excreted by the bacterium (as in food poisoning caused by *Clostridium botulinum*), (2) the elaboration of an endotoxin, a harmful chemical substance that is liberated only after disintegration of the microorganism (as in typhoid, caused by *Salmonella typhi*), or (3) the induction of sensitivity within the host to antigenic properties of the bacterial organism (as in tuberculosis, after sensitization to *Mycobacterium tuberculosis*).

Fungi and other parasites. Diseases caused by fungi and parasites are relatively uncommon in developed countries. Fungal infections, also known as mycotic infections, may affect the skin surfaces or the internal organs of the body. The superficial mycotic infections are generally not serious and include such well-known disorders as athlete's foot (tinea pedis), caused by the dermatophyte *Trichophyton*. Deep mycotic infections such as histoplasmosis and candidiasis are potentially life-threatening.

Other parasites that attack humans range in size from unicellular organisms such as *Entamoeba histolytica* to such multicellular forms as tapeworms and roundworms. Most parasitic infestations are encountered in the less-developed areas of the world where sanitation is not optimal. Indeed, parasitic infestations constitute major causes of death in regions of Central and South America, Africa, India, and Asia. (For additional information about diseases of biotic origin, see INFECTIOUS DISEASES.)

ABNORMAL GROWTH OF CELLS

Normal and abnormal cell growth. *Cell growth inhibition.* The growth of cells in the body is a closely controlled function, which, together with limited and regulated expression of various genes, gives rise to the many different tissues that constitute the whole organism. For the most part, control of cell growth persists throughout life except for episodic instances such as healing of an injured tissue. In this situation the growth of a localized group of cells is accelerated to reconstitute the tissue to its previous state of normal structure and function, following which tightly regulated growth resumes. Such areas of increased cell growth are referred to as hyperplasias; they consist of expanded numbers of normal-appearing cells and, depending on the duration of growth, can result in an enlargement of tissues and organs. In general, hyperplasias arise to meet special needs of the body and subside once these needs are met. Hyperplasias are the result of the sustained impact over time of stimulatory influences together with a loss of growth-inhibitory factors that are normally found within or around cells. As long as the loss of inhibition of cell growth is temporary, the capacity for enhanced cell proliferation when necessary has obvious advantages. However, if cells permanently lose their ability to respond to growth-inhibitory factors, their growth becomes irrepressible, and cancer may result.

Hyperplasia

Neoplasms: malignant and benign tumours. Diseases arising from uncontrolled cell growth and behaviour collectively constitute the second most common cause of human death (the most common cause being heart disease). Cancers, the most important form of abnormal growth and behaviour, were responsible for approximately 538,000 deaths, or almost one-fourth of all deaths, in the United States in 1994. The significance of this incidence is placed in proper perspective by a consideration of the following facts. While cancer arises at all stages of life, its incidence (number of cases) increases with age, reaching a peak between 55 and 74 years. This fact, together with the increasing longevity of the general population and improved diagnostic modalities that enable clinicians to detect cancers with greater frequency, tempers the notion that the incidence of cancer is increasing.

In addition to cancers—malignant tumours that may eventually kill the host—there are benign tumours that rarely produce serious disease. The two types of tumours are collectively referred to as neoplasms (new growths), and their study is known as oncology. Tumours are referred to as malignant or benign based on the structural and functional properties of their component cells and their biological behaviour. The cells and tissues of malignant tumours differ from the tissues from which they arise. They exhibit more rapid growth and altered structure and function, including stimulation of new blood vessel growth (angiogenesis) and a capacity to invade adjacent normal tissues, enter the blood vascular system, and spread (metastasize) to distant sites. The properties of malignant tumour cells serve to enhance and support their proliferation and extension throughout the body tissues and organs, eventually leading to death of the host. In contrast, the cells and tissues of benign tumours tend to grow more slowly and in general closely resemble their normal tissues of origin. When the structure and function of benign tumour cells are morphologically and functionally indistinguishable from those of normal cells, their growth as a tumour mass is the sole feature indicative of their neoplastic nature. It is hoped that a greater understanding of malignant cell growth and behaviour will lead to the development of novel cancer therapies based on tumour cell biology that will complement or replace the current treatments of surgical extirpation (complete excision), chemotherapy, and radiation.

Growth characteristics of tumours

Characteristics of cancer. *Epidemiology.* Epidemiological studies of the worldwide incidence of cancers have identified striking differences among countries and population groups. For example, the incidence of and death rates for skin cancer are much higher in Australia and New Zealand than in the Scandinavian countries—presumably because of the marked differences between these two regions in total annual hours of exposure to sunlight. The importance of environmental influences is highlighted by comparing the incidence of and death rates for cancers among populations in different geographic regions. For example, prostate and colon cancer rates in Japanese persons living in Japan differ from the rates in Japanese persons who have emigrated to the United States, the rates of their offspring born in California, and the rates of long-term white residents of that state. These rates are much lower among Japanese living in Japan than they are in white Californians. However, the rates for each type of tumour among first-generation Japanese immigrants are intermediate between the rates in Japan and those in California, suggesting that environmental and cultural factors may play a more important role than genetic ones.

The role of genetics. The irreversibility of the structural and behavioral changes of cancer cells has long been recognized and has favoured the postulate that they are probably due to permanent genetic alterations. This postulate remained speculative until the discovery in 1979 that oncogenes (cancer-causing genes) are derived from proto-oncogenes (normal growth-regulatory cellular genes). When proto-oncogenes become mutated or deregulated, they are converted to oncogenes, which are capable of causing the malignant transformation of cells, including those of humans. Cellular proto-oncogenes code for proteins involved in cell regulation, such as growth factors, their receptors, and transmembrane signal transducers. Thus, changes in the structure of proto-oncogenes and

their conversion to oncogenes results in the synthesis of abnormal proteins that are incapable of carrying out their usual growth-regulatory functions. In identifying the genes involved in the development of cancer, researchers discovered a group of cellular genes—tumour-suppressor, or suppressor, genes—whose protein products normally negatively regulate cell growth by suppressing cell proliferation, thus counterbalancing the growth-stimulatory effects of proteins synthesized by proto-oncogenes. Genetic analyses of various animal and human cancers have demonstrated that, in the majority, alterations of oncogenes and suppressor genes were often simultaneously present. These analyses suggest that multiple genetic alterations involving growth-stimulatory and growth-inhibitory genes are required for the induction of malignancy. Such discoveries have ushered in a new era in cancer biology and may well lead to the eventual control, cure, and prevention of malignant diseases.

Heredity and environment. The many causes of cancer include intrinsic factors, such as heredity, and extrinsic factors, such as environment and lifestyle. Hereditary causes of cancer are less common and are due to the inheritance of a single mutant gene that greatly increases the risk of developing a malignant tumour. Such cancers include (1) a childhood tumour of the eye, retinoblastoma, and a bone tumour, osteosarcoma, both of which involve the loss of a tumour suppressor gene, and (2) familial adenomatous polyposis, in which all patients develop colon cancer by age 50. The most common types of cancer that occur sporadically, such as cancers of the breast, ovary, colon, and pancreas, also have been documented to occur in familial forms. The children in such families appear to have a two- to threefold increased risk of developing a particular tumour, but the transmission pattern is unclear. A still rarer hereditary cause of cancer is an inherited deficiency in the ability to repair DNA. Patients with this defect (known as xeroderma pigmentosum) are particularly sensitive to sunlight and develop skin cancer during early adolescence because of unrepaired mutations induced by ultraviolet (UV) light.

Familial types of cancer

Although the environment contains many agents that can cause cancer in humans, the extent to which they contribute to the human disease is often difficult to assess. For example, the link between tobacco smoking and lung cancer is clear; however, little is known about the cause of cancer of the prostate, the most common form of cancer in males, despite the fact that many factors—including age, race, male hormone, increased consumption of dietary fat, and a genetic basis—have been implicated.

Three categories of carcinogens (chemical or physical agents that mutate DNA) that induce cancer in experimental animals and humans have been identified in the environment: (1) chemicals, (2) radiant energy, and (3) oncogenic viruses.

Carcinogenic agents. *Chemicals.* Chemicals capable of causing cancer arise from a variety of sources. These include certain synthetic chemicals used in industry, some natural compounds formed during the curing and burning of tobacco, compounds formed during the cooking of meat, and chemicals present in certain plants and molds. Two categories have been identified, those capable of causing DNA damage and mutations directly (genotoxic, or direct-acting, carcinogens) and those that require prior metabolic activation by cells of the host to be converted to mutagens (epigenic, or indirect-acting, carcinogens). In the industrial countries much progress has been made in significantly decreasing and preventing exposure to chemical carcinogens in the workplace. However, exposure to carcinogens as a consequence of cultural practices, such as tobacco smoking and the cooking and consumption of meats, is difficult if not impossible to control or eradicate.

Radiant energy. Sustained exposure to two forms of radiant energy—namely, UV light and ionizing radiation—is carcinogenic for humans. Repeated and sustained exposure to UV rays emanating from the Sun causes mutations of DNA that ultimately are capable of inducing three different types of skin cancer. As one would expect, the incidence of UV-induced skin cancer is high among farmers, sailors, and sunbathing enthusiasts. The degree of risk depends on the extent of exposure and the amount of melanin pigment in the skin, which absorbs UV rays. Dark-skinned individuals are protected by the high content of melanin in their skin; in contrast, fair-skinned persons and albinos have very little or no protective melanin pigment in their skin.

UV light and ionizing radiation

The carcinogenic effects of ionizing radiation first became apparent from the results of inappropriate exposure of early uranium ore miners and of physicians who first used X-ray machines for diagnostic purposes and were unaware of the health hazards. The devastating complications that resulted are rare today because of stricter indications for the use of radiation therapy, careful focusing of radiation beams, and effective shielding of adjacent normal tissues. However, the risks of exposure to ionizing radiation have been reemphasized from time to time by the appearance of neoplastic disease following radiation therapy and following the release of enormous amounts of radiation into the environment, as occurred from atomic bombing of Hiroshima and Nagasaki in Japan and the accident at the Chernobyl nuclear power station in Ukraine.

Reactive forms of carcinogenic chemicals and, in the case of ionizing radiation, reactive forms of oxygen damage DNA directly. If repair of damaged DNA is slow, error-prone, or not accomplished at all and cell replication occurs, the damage is amplified and becomes a permanent (fixed) mutation.

Viruses. In recent years certain DNA viruses have been strongly implicated as causal agents for a variety of cancers in humans. These include human papillomavirus (HPV) as a cause of genital cancers in both sexes worldwide, the Epstein-Barr virus (EBV) for childhood lymphoma in Africa and cancer of the nose and throat in Asia and Africa, and the hepatitis viruses B and C that cause liver cancer worldwide with the highest incidence in Asia and Africa. However, at present only one type of human cancer, the rare adult T-cell leukemia, has been solidly linked to infection with an RNA virus, the human T-cell leukemia virus (HTLV-1). While much experimental and clinical evidence supports the carcinogenic role of the above-mentioned viruses in humans, additional research suggests that other factors also may be required. Thus far, oncogenic viruses have not been shown to induce DNA mutations directly in human cells; rather, their contribution seems to lie in promoting and hastening the process of mutation. (For greater detail on how viruses contribute to the induction of cancer, see the articles CANCER and VIRUSES.)

DISEASES OF METABOLIC-ENDOCRINE ORIGIN

The term metabolism encompasses all the chemical reactions vital to the growth and maintenance of the body. Defects in metabolism are found in almost every disease condition. Most are secondary; *i.e.,* they result from some other basic disorder (infection, kidney disease, or heart disease, for example). In a few primary metabolic disorders, small genetic mutations lead to structural alterations of specific proteins that disrupt protein function and are responsible for the disease state. At this point, another group of primary metabolic disorders—those associated with hormonal defects—will be touched on.

Hormones are large organic molecules secreted in small amounts by specific cells in the various endocrine (ductless) glands. These secretions are carried by the blood to distant sites (target organs), where they bind to specific receptors on target cells and act to regulate specific chemical reactions.

Cause of endocrine disease

All endocrine disease stems from either an overproduction (hyperfunction) or underproduction (hypofunction) of some hormone-secreting endocrine gland. There are relatively few causes of hormone overproduction. In general, overproduction results from hyperplasia, an increase in the number of cells (in this case, hormone-secreting cells) in a specific endocrine gland. It can also be caused by neoplasia, the growth of a tumour in an endocrine gland. Although most endocrine tumours are benign, the resulting hypersecretion of hormone can have far-reaching effects. For example, the pituitary gland, tucked into the base of the skull, produces many hormones that have far-ranging

effects, mostly controlling the function of the other endocrine glands, such as the thyroid, adrenals, ovaries, and testes. Acromegaly, characterized by the enlargement of many skeletal parts, is a rare endocrine disease caused by excess secretion of pituitary growth hormone in the adult.

Underproduction of hormone is most often the result of destruction of hormone-secreting cells. This destruction may be caused by infection, infarction (tissue death due to loss of blood supply), or obliteration of endocrine glands by cancer. Underproduction of hormone also may result from failure of the gland to undergo normal fetal development, or it may be a feature of an autoimmune disease (as in juvenile diabetes mellitus).

Treatment of endocrine disease involves either hormone supplementation, in the case of hypofunction, or, in cases of hyperfunction, destruction of endocrine gland tissue by surgery or radiation (see ENDOCRINE SYSTEMS).

DISEASES OF NUTRITION

Diseases of nutrition include the effects of undernutrition, prevalent in less-developed areas but present even in affluent societies, and the effects of nutritional excess.

Diseases of nutritional excess. Obesity, perhaps the most important nutritional disease in the United States and Europe, results usually from excessive caloric intake, although emotional, genetic, and endocrine factors may be present.

Obesity predisposes one toward several serious disorders, including a state of chronic oxygen deficiency called the hypoventilation syndrome; high blood pressure; and atherosclerosis, a degenerative condition of the blood vessels that is discussed further below.

Excessive intake of certain vitamins, especially vitamins A and D, can also produce disease. Vitamins A and D are both fat-soluble and tend to accumulate to toxic levels in the bodily tissues when taken in excessive quantities. Vitamin C and the B vitamins, soluble in water, are more easily metabolized or excreted and, therefore, rarely accumulate to toxic levels.

Types of nutritional deficiency

Diseases of nutritional deficiency. Nutritional deficiencies may take the form of inadequacies of (1) total caloric intake, (2) protein intake, or (3) certain essential nutrients such as the vitamins and, more rarely, specific amino acids (components of proteins) and fatty acids.

Protein-calorie malnutrition remains prevalent in certain areas. It has been estimated that about two-thirds of the world's population has less than enough food to eat. Not only is the quantity inadequate but the quality of the food is nutritionally deficient and usually lacks protein. In deprived areas malnutrition has its greatest impact on the young. Deaths from protein-calorie malnutrition result from the failure of the child to thrive, with progressive weight loss and weakness, which in turn can lead to infection and disease, usually some form of gastrointestinal bacterial or parasitic disorder. In other circumstances adequate calories may be available, but a deficiency of protein induces a disorder known as kwashiorkor.

Vitamin deficiencies, the most important forms of selective malnutrition, may arise in a variety of ways, the most common and the most important being an improper, inadequate diet. When the total caloric intake is inadequate, vitamin deficiencies may also occur, but in these circumstances the more profound lack of calories and proteins masks the lack of vitamins.

Secondary vitamin deficiencies

Vitamin deficiencies may also be encountered despite a diet that is apparently adequate nutritionally. One source of such a deficiency, called secondary, is interference with absorption of the vitamin. Pernicious anemia is a classic example of this phenomenon. This disorder results from an autoimmune response to intrinsic factor, a substance normally found in the stomach lining with which vitamin B_{12} must form a complex to be absorbed. (Vitamin B_{12} is necessary for red cells to form properly.) The basis of pernicious anemia, then, is a lack of absorption of vitamin B_{12}. The absence of certain digestive enzymes, as is found in pancreatic disease, can lead to the inability to digest and absorb fats and the fat-soluble vitamins (A, D, E, and K). Impaired uptake of vitamins may be encountered in gastrointestinal diseases. Some of these diseases reduce the absorptive function of the bowel. Similarly, diseases associated with severe, prolonged vomiting may interfere with adequate absorption.

Avitaminosis (vitamin lack) may be encountered when there are increased losses of vitamins such as occur with chronic severe diarrhea or excessive sweating or when there are increased requirements for vitamins during periods of rapid growth, especially during childhood and pregnancy. Fever and the endocrine disorder hyperthyroidism are two additional examples of conditions that require higher than the usual levels of vitamin intake. Unless the diet is adjusted to the increased requirements, deficiencies may develop. Lastly, artificial manipulation of the body and its natural metabolic pathways, as by certain surgical procedures or the administration of various drugs, can lead to avitaminoses. (Diseases involving deficiencies of particular vitamins are discussed in NUTRITION: *Deficiency diseases: Vitamins*.)

DISEASES OF NEUROPSYCHIATRIC ORIGIN

Diseases of neuropsychiatric origin afflict large segments of the population. For example, a total of about 2.8 million persons in the United States suffer from three major psychiatric diseases—schizophrenia, major depression, and mania—and three major neurological disorders—Alzheimer's disease, Huntington's chorea, and Parkinson's disease. These six conditions will be briefly reviewed here. More in-depth coverage is found in the articles MENTAL DISORDERS AND THEIR TREATMENT and NERVES AND NERVOUS SYSTEMS.

The key function of the nervous system is to collect information about the body and its external environment, process the information, and coordinate the body's responses to that information. This complex function depends on each nerve cell (neuron) receiving signals from other neurons and transmitting this input to still other neurons. This critical input and output of communication (signaling) between neurons is mediated by chemical transmitter molecules (neurotransmitters). Neurotransmitters are synthesized by nerve cells and released from one cell to another across a narrow gap between the two neurons known as the synapse. Eight different major neurotransmitters and a large number of neuropeptide molecules (which serve to modulate the effects of neurotransmitters) have been identified. Different types of nerve cells respond to different neurotransmitters and neuropeptides. Chemical signaling between nerve cells is rapid and precise and can occur over long distances. The precision is due to receptor molecules, which are activated following their recognition and binding of specific neurotransmitters. In some types of nerves the synapses do not possess receptors, in which case interneuronal communication is achieved by electrical transmission. In many neuropsychiatric diseases alterations in the levels of transmitter substances appear to play a major role in pathogenesis.

Communication within the nervous system

Psychiatric diseases. Mental illnesses affect the very fabric of human nature, robbing it of its various facets of personality, purposeful behaviour, abstract thinking, creativity, emotion, and mood. Those suffering from mental disorders exhibit a spectrum of symptoms depending on the severity of their disease. These diseases include obsessive-compulsive personality disorder, dementia, schizophrenia, major depression, and manic disorders.

Schizophrenia

Schizophrenia in its severe form is a catastrophic mental illness that begins in adolescence or early adult life. It is relatively common, occurring in about 1 percent of the general population worldwide. Because the incidence of schizophrenia among parents, children, and siblings of patients with the disease is increased to 15 percent, it is believed that heredity plays an important role in the genesis of the disease. However, other studies suggest that nongenetic factors are also influential. The biochemical basis of the disease may be an excess of the neurotransmitter substance dopamine, as high levels of dopamine and its metabolites, as well as increased dopamine receptors, are found in the brains of persons with schizophrenia. Further evidence for this hypothesis is that the drugs most effective in treating the disease are those that have a high capacity to block dopamine receptors.

Pathological disturbances of mood, ranging from severe depression to manic behaviour, are common forms of mental illnesses. Severe depression is characterized by despondency, diminished interest in most or all activities, weight fluctuation not due to dieting, disruption in sleep patterns, psychomotor agitation or retardation, feelings of worthlessness, excessive quiet, and recurrent thoughts of death or suicide. Manic behaviour involves a period in which an expansive, elevated, or irritable mood persists abnormally. During this episode symptoms such as increased talkativeness, distractibility, decreased need for sleep, inflated self-esteem, and excessive involvement in pleasurable yet risky activities may be present. Major depression is associated with decreased brain levels of the neurotransmitters norepinephrine and serotonin, and the most effective therapy consists of drugs that inhibit the breakdown of these compounds. The neurochemical alterations in mania are less clearly understood, but it is well established that drugs effective in the treatment of mania are those that antagonize dopamine and serotonin. The mechanism responsible for the therapeutic efficacy of lithium for the treatment of mania is not yet clear. Although mood disorders have a familial background, the evidence for a genetic component is not convincing.

Neurological diseases. The three neurological diseases considered in this section—Alzheimer's disease, Huntington's chorea, and Parkinson's disease—are age-related, and to varying degrees they manifest as deterioration of mental function that involves the loss of memory and of acquired intellectual skills. This deterioration is referred to as dementia. Because dementia can result from many causes, other features of each disease must be present before a definitive diagnosis can be made.

Alzheimer's disease. Alzheimer's disease is the most common form of dementia, being responsible for about two-thirds of the cases of dementia in patients over 60 years of age. Women are affected twice as often as men. More rarely there are familial forms of the disease that have an early onset affecting individuals in the fourth and fifth decades of life. Alzheimer's disease is insidious in onset. Early manifestations include memory loss, temporary confusion, restlessness, poor judgment, and lethargy. A failure to retain new information and a deterioration of social relationships often ensue. In some patients paranoia and delusions are the first symptoms of the disease. Whatever the onset, the last stages are characterized by intellectual vacuity and loss of control over all body functions.

The brains of patients with Alzheimer's disease are characterized by the loss of neurons, which, as the disease progresses, becomes severe and leads to decreased brain size and weight. Because nerve cells synthesize the neurotransmitters necessary for interneuronal communication, it is not surprising that Alzheimer's disease is associated with diminished levels of neurotransmitters, including acetylcholine, norepinephrine, and serotonin, as well as modulatory neuropeptide molecules that transmit signals between nerve cells. Two other characteristic tissue lesions found in the cerebral cortex of patients with Alzheimer's disease are neuritic plaques and neurofibrillary tangles. Neuritic plaques are deposits of neuron fragments surrounding a core of amyloid β-protein. Neurofibrillary tangles are twisted fibres of the protein tau found within neurons.

A variety of genetic factors have been identified in the different forms of Alzheimer's disease. The rare cases of the early familial forms of the disease are linked to three different genetic defects found on three different chromosomes—chromosomes 1, 14, and 21. Another gene on chromosome 19 is believed to play a part in the more common late-onset cases. The gene on chromosome 21 was the first to be identified. (This finding is significant because an abnormality in chromosome 21—an extra copy—is found in patients with Down syndrome, virtually all of whom develop Alzheimer's disease if they live to age 35.) The defective gene on chromosome 21 normally codes for amyloid precursor protein. A defect in this gene is thought to result in abnormal cleavage of the protein that increases the production and deposition of amyloid β-protein. This gene, however, is linked to only 2 to 3 percent of all early familial cases of the disease. The majority of patients with early-onset disease—70 to 80 percent—have the genetic mutation on chromosome 14, and another group of patients have a defective gene on chromosome 1. The gene on chromosome 19 codes for apolipoprotein E, a protein involved in cholesterol transport and metabolism. Three forms, or alleles, of the gene exist. The presence of one form—ApoE4—in an individual's genome seems to increase the deposition of amyloid β-protein in the brain and may also increase the number of neurofibrillary tangles.

Huntington's chorea. Huntington's chorea occurs at the rate of about 5 per 100,000 individuals. It affects both sexes equally and usually becomes manifest in the fourth decade of life. The disorder is characterized by uncontrolled movements (chorea), dementia, and death within 20 years after onset. The symptoms worsen until the patient becomes totally incapacitated and bedridden. Huntington's chorea is a hereditary disease passed on as an autosomal dominant trait (see above *Diseases of genetic origin*). Because of its highly regular familial inheritance, the disease is often traceable to the original carriers who introduced the defective gene. For example, British immigrants to colonial America in the 17th century are believed to be responsible for almost all cases of Huntington's chorea in the eastern United States, and an English sailor is thought to have introduced the defective gene into Venezuela almost 200 years ago. The recent localization of the Huntington's chorea gene to chromosome 4 and its cloning will allow identification of the gene product, insight into the mechanism responsible for the disease, and perhaps effective treatment. It will also permit the disease to be diagnosed in fetuses as well as in children before the onset of symptoms.

Parkinson's disease. Parkinson's disease is a motor disorder characterized by the onset of a "pill rolling" rhythmic tremor, muscle rigidity, difficulty and slowness in movement, and stooped posture. As the disease progresses, the face of the patient becomes expressionless, the rate of swallowing is reduced, leading to drooling, and depression and dementia increase. The prevalence of Parkinson's disease is estimated to be about 187 per 100,000 persons in the general population, with about 20 new cases per 100,000 appearing each year. Men are slightly more affected than women, and there are no apparent racial differences. The disease appears typically in the sixth and seventh decades, although occasionally it can begin as early as the third decade. While the majority of cases of Parkinson's disease are of unknown cause, the disease was linked to a complication of encephalitis that developed following the worldwide outbreak of influenza during World War I and, more recently, to an episode of use by intravenous drug abusers of a synthetic drug similar to heroin. A marked decrease in the level of dopamine, a major neurotransmitter, has been noted in the brains of patients with Parkinson's disease, and this change has been attributed to the loss of so-called dopaminergic neurons that normally synthesize and use dopamine to communicate with other neurons in parts of the brain that regulate motor function. This information has opened a new approach to the treatment of the disease—namely, administration of the metabolic precursor to dopamine (L-dopa) that can be converted by the body to dopamine. Although initially beneficial in causing a significant remission of symptoms, L-dopa frequently is effective for only 5 to 10 years, and serious side effects accompany treatment. Cotreatment with an inhibitor of the enzyme that breaks down L-dopa and thus allows the substance to remain in the brain longer has yielded an effective therapy, which allows many patients to live reasonably normal lives. Treatment of the disease with fetal tissue transplants, while raising ethical and legal issues, also has shown promise in alleviating symptoms. Nevertheless, although treatment may slow the progress of the disease, it does not alter its course. This suggests that factors other than variation in neurotransmitter levels are responsible for the disease.

Treatment of Parkinson's disease

DISEASES OF SENESCENCE

The process of aging begins at the time of conception. Throughout life the body undergoes a series of changes

that can be considered as manifestations of aging. During the first half of life these changes are generally referred to as maturation, during the last half of life as progressive senescence. Visual acuity, sensitivity of hearing, and muscular vigour begin to deteriorate after the third decade of life. These changes, although they may begin at different ages and progress at differing rates, are universal among all individuals and must therefore be considered as the normal aging process. A critical question remains unanswered concerning the cause of the intrinsic retrogressive changes in cell and organ structure and function that occur throughout the aging process. Are these changes genetically determined, or are they a result of accumulated sublethal injuries that the cell sustains from exposure to noxious environmental factors over time? Or perhaps both elements act in concert to effect the changes that occur as life progresses.

It is extremely difficult to draw a sharp line between the deleterious effects of normal aging and the deleterious effects of the diseases of aging. The diseases most commonly manifested in the elderly are disorders of the heart, blood vessels, and joints. The heart disease of the elderly is related to the generalized vascular disease known as arteriosclerosis, which frequently attacks the major coronary arteries of the heart. Arteriosclerosis and arthritis will therefore be briefly touched upon here. More extended discussions may be found in CIRCULATION AND CIRCULATORY SYSTEMS: *Cardiovascular system diseases and disorders* and in SUPPORTIVE AND CONNECTIVE TISSUES: *Joint diseases and injuries*. These problems and other aspects of aging are also considered in GROWTH AND DEVELOPMENT: *Human aging*.

Arteriosclerosis is not a specific disease. The term is applied to all diseases that cause hardening of the arteries. Several minor processes can induce hardening of the arteries, but the overwhelming preponderance of cases of arteriosclerosis are caused by atherosclerosis. This disorder, which eventually affects all individuals to varying degrees, begins relatively early in life in most persons. There are great variations, however, in the severity of this disease among individuals and among racial, national, and ethnic populations. These differences depend on the presence or absence of risk factors such as diet, hypertension, tobacco smoking, diabetes, obesity, family history, and stress.

Atherosclerosis

Atherosclerosis is characterized by the deposition of fats (cholesterol and other complex lipids) in the linings (intima) of the arteries. It is accompanied by cell injury, cell death, and scarring and sometimes produces total obstruction of an artery. Atherosclerosis has a predilection for the aorta, the major artery of the body, and the arteries of the heart, brain, and legs. Atherosclerosis of the arteries of the heart (the coronaries) causes myocardial infarction, otherwise known as heart attack.

When atherosclerosis narrows but does not totally block the coronary arteries, the heart also is injured by lack of adequate blood supply and nutrition and becomes progressively smaller and weaker; even though this disease is not as life-threatening as a heart attack, it nonetheless frequently causes heart failure, an inability of the heart to deliver an adequate supply of blood to the tissues. Atherosclerosis of the arteries of the brain is the usual cause of stroke. When the arteries to the legs become affected in this way, gangrene may develop.

Arthritis, probably the second most common and distressing disease among the elderly, is a disease of the joints. It causes considerable pain, discomfort, and lack of mobility and so makes life burdensome. Moreover, arthritic individuals are more subject to other illnesses. Degenerative arthritis (osteoarthritis) is common to all elderly people to a lesser or greater degree. Osteoarthritis usually begins in the fourth decade of life and slowly progresses with increasing age. Coinciding with the characteristic degeneration of the joints are changes involving the bone itself. The bone of elderly persons is known to be less dense and more brittle; it tends, therefore, to fracture more easily. It also heals with greater difficulty.

There are many subtle changes that occur with the normal aging process. These may include degenerative changes in the brain, leading to impaired mental ability and even senility. As this damage is usually accompanied by atherosclerosis of the arteries of the brain, it is difficult to know how much of the change is the result of impaired blood flow and how much is related to normal aging. Finally, but of no less significance, is the general decline in the body's ability to defend itself against disease. Thus elderly persons are more susceptible to infections, trauma, and a number of other bodily defects. Simple, uncomplicated pneumonia, which might be easily tolerated by the young, healthy adult, may be fatal for an elderly, weakened person.

Classifications of diseases

Value of disease classification

Classifications of diseases become extremely important in the compilation of statistics on causes of illness (morbidity) and causes of death (mortality). It is obviously important to know what kinds of illness and disease are prevalent in an area and how these prevalence rates vary with time. Classifying diseases made it apparent, for example, that the frequency of lung cancer was entering a period of alarming increase in the mid-20th century. Once a rare form of cancer, it had become the single most important form of cancer in males. With this knowledge a search was instituted for possible causes of this increased prevalence. It was concluded that the occurrence of lung cancer was closely associated with cigarette smoking. Classification of disease had helped to ferret out an important, frequently causal, relationship.

The most widely used classifications of disease are (1) topographic, by bodily region or system, (2) anatomic, by organ or tissue, (3) physiological, by function or effect, (4) pathological, by the nature of the disease process, (5) etiologic (causal), (6) juristic, by speed of advent of death, (7) epidemiological, and (8) statistical. Any single disease may fall within several of these classifications.

In the topographic classification, diseases are subdivided into such categories as gastrointestinal disease, vascular disease, abdominal disease, and chest disease. Various specializations within medicine follow such topographic or systemic divisions, so that there are physicians who are essentially vascular surgeons, for example, or clinicians who are specialized in gastrointestinal disease. Similarly, some physicians have become specialized in chest disease and concentrate principally on diseases of the heart and lungs.

In the anatomic classification, disease is categorized by the specific organ or tissue affected; hence, heart disease, liver disease, and lung disease. Medical specialties such as cardiology are restricted to diseases of a single organ, in this case the heart. Such a classification has its greatest use in identifying the various kinds of disease that affect a particular organ. The heart is a good example to consider. By the segregation of cardiac disease it has been made apparent that heart disease is now the most important cause of death in the United States and in most other industrialized nations. Moreover, it has become apparent that disease caused by atherosclerosis of the coronary arteries is by far the most important form of heart disease. In making a diagnosis of cardiac disease in an elderly patient, the cardiologist must first determine whether this disease of the coronary arteries is responsible for the heart's failure to function normally.

Physiological classification of disease

The physiological classification of disease is based on the underlying functional derangement produced by a specific disorder. Included in this classification are such designations as respiratory and metabolic disease. Respiratory diseases are those that interfere with the intake and expulsion of air and the exchange of oxygen for carbon dioxide in the lungs. Metabolic diseases are those in which disturbances of the body's chemical processes are a basic feature. Diabetes and gout are examples.

The pathological classification of disease considers the nature of the disease process. Neoplastic and inflammatory disease are examples. Neoplastic disease includes the whole range of tumours, particularly cancers, and their effect on human beings.

The etiologic classification of disease is based on the cause, when known. This classification is particularly important and useful in the consideration of biotic disease.

On this basis disease might be classified as staphylococcal or rickettsial or fungal, to cite only a few instances. It is important to know, for example, what kinds of disease staphylococci produce in human beings. It is well known that they cause skin infections and pneumonia, but it is also important to note how often they cause meningitis, abscesses in the liver, and kidney infections. The sexually transmitted diseases syphilis and gonorrhea are further examples of diseases classified by etiology.

The juristic basis of the classification of disease is concerned with the legal circumstances in which death occurs. It is principally involved with sudden death, the cause of which is not clearly evident. Thus, on a juristic basis some deaths and diseases are classified as medical-legal and fall within the jurisdiction of coroners and medical examiners. A person living alone is found dead in bed—dead of natural causes or killed? Had the person who dropped dead on the street been given some poison that took a short time to act? Much less dramatic, but perhaps more common, are disease and death caused by exposure of the individual to some unrecognized danger to health in working or living conditions. Could the illness or disease be attributable to fumes or dusts in a factory? These are examples of the many types of disease and death that fall properly in this classification.

Epidemiological classification

The epidemiological classification of disease deals with the incidence, distribution, and control of disorders in a population. To use the example of typhoid, a disease spread through contaminated food and water, it first becomes important to establish that the disease observed is truly caused by *Salmonella typhi,* the typhoid organism. Once the diagnosis is established, it is obviously important to know the number of cases, whether the cases were scattered over the course of a year or occurred within a short period, and what the geographic distribution is. It is critically important that the precise address and activities of the patients be established. Two widely separated locations within the same city might be found to have clusters of cases of typhoid all arising virtually simultaneously. It might be found that each of these clusters revolved about a family unit including cousins, grandparents, aunts and uncles, and friends, suggesting that in some way personal relationships might be important. Further investigation might disclose that all the infected persons had dined at one time or at short intervals in a specific home. It might further be found that the person who had prepared the meal had recently visited some rural area and had suffered a mild attack of the disease and was now spreading it to family and friends by unknowing contamination of food. This hypothetical case suggests the importance of the etiologic, as well as the epidemiological, classification of disease.

Epidemiology is one of the important sciences in the study of nutritional and biotic diseases around the world. The United Nations supports, in part, the World Health Organization, whose chief function is the worldwide investigation of the distribution of disease. In the course of this investigation, many observations have been made that help to explain the cause and provide approaches to the control of many diseases.

Statistical classification

The statistical basis of classification of disease employs analysis of the incidence (the numbers of new cases of a specific disease that occur during a certain period) and the prevalence rate (number of cases of a disease in existence at a certain time) of diseases. If, for example, a disease has an incidence rate of 100 cases per year in a given locale and, on the average, the affected persons live three years with the disease, it is obvious that the prevalence of the disease is 300. Statistical classification is an additional important tool in the study of possible causes of disease. These studies, as well as epidemiological, nutritional, and pathological analyses, have made it clear, for example, that diet is an important consideration in the possible causation of atherosclerosis. The statistical analyses drew attention to the role of high levels of fats and carbohydrates in the diet in the possible causation of atherosclerosis. The analyses further drew attention to the fact that certain populations that do not eat large quantities of animal fats and subsist largely on vegetable oils and fish have a much lower incidence of atherosclerosis. Thus, statistical surveys are of great importance in the study of human disease.

(S.L.R./J.H.Ro./D.G.Sc.)

DISEASES OF ANIMALS

Concern with diseases that afflict animals dates from the earliest human contacts with animals and is reflected in early views of religion and magic. Diseases of animals remain a concern principally because of the economic losses they cause and the possible transmission of the causative agents to humans. The branch of medicine called veterinary medicine deals with the study, prevention, and treatment of diseases not only in domesticated animals but also in wild animals and in animals used in scientific research. The prevention, control, and eradication of diseases of economically important animals are agricultural concerns. Programs for the control of diseases communicable from animals to man, called zoonoses, especially those in pets and in wildlife, are closely related to human health. Further, the diseases of animals are of increasing importance, for a primary public-health problem throughout the world is animal-protein deficiency in the diet of humans. Indeed, both the United Nations Food and Agricultural Organization (FAO) and the World Health Organization (WHO) have been attempting to solve the problem of protein deficits in a world whose human population is rapidly expanding.

General considerations

HISTORICAL BACKGROUND

Parallel development with human medicine

Historical evidence, like that from currently developing nations, indicates that veterinary medicine originally developed in response to the needs of pastoral and agricultural man along with human medicine. It seems likely that a veterinary profession existed throughout a large area of Africa and Asia from at least 2000 BC. Ancient Egyptian literature includes monographs on both animal and human diseases. Evidence of the parallel development of human and veterinary medicine is found in the writings of Hippocrates on medicine and of Aristotle, who described the symptomatology and therapy of the diseases of animals, including man. Early Greek scholars, noting the similarities of medical problems among the many animal species, taught both human and veterinary medicine. In the late 4th century BC, Alexander the Great designed programs involving the study of animals, and medical writings of the Romans show that some of the most important early observations on the natural history of disease were made by men who wrote chiefly about agriculture, particularly the aspect involving domesticated animals.

Most of the earliest suggestions of relationships between human health and animal diseases were part of folklore, magic, or religious practice. The Hindu's concern for the well-being of animals, for example, originated in his belief in reincarnation. From the pre-Christian Era to about 1500, the distinctions between the practices of human and veterinary medicine were not clear-cut; this was especially true in the fields of obstetrics and orthopedics, in which animal doctors in rural areas often delivered babies and set human-bone fractures. It was realized, however, that training in one field was inadequate for practicing in the other, and the two fields were separated.

Veterinary literature from the civilizations of Greece and Rome contains reference to "herd factors" in disease; contagion within groups of animals kept together, therefore, was recognized, and both quarantine and slaughter were used to control outbreaks of livestock diseases. Rinderpest (cattle plague) was the most important livestock disease from the 5th century until control methods were developed. Serious outbreaks of the disease prompted the

founding of the first veterinary college (École Nationale Vétérinaire), in Lyon, France, in 1762. Many aspects of animal diseases are best understood in terms of population or herd phenomena; for example, herds of livestock, rather than individual animals, are vaccinated against specific diseases, and housing, nutrition, and breeding practices are related to the likelihood of illness in the herd.

Pasteur's impact

The work of Pasteur was of fundamental significance to general medicine and to agriculture. Veterinarians became concerned with foods of animal origin after the discovery of microorganisms and their identification with diseases in man and other animals. Efforts were directed toward protecting humans from diseases of animal origin, primarily those transmitted through meat or dairy products. Modern principles of food hygiene, first established for the dairy and meat-packing industries in the 19th and early 20th centuries, have been generally applied to other food-related industries. The veterinary profession, especially in Europe, assumed a major role in early food-hygiene programs.

Since World War II, the eradication of animal diseases, rather than their control, has become increasingly important, and conducting basic research, combatting zoonoses, and contributing to man's food supply have become indispensable services of veterinary medicine.

IMPORTANCE

Economic importance. About 50 percent of the world's population suffers from chronic malnutrition and hunger. Inadequate diet claims many thousands of lives each day. When the lack of adequate food to meet present needs for an estimated world population of more than 4,600,000,000 in the 1980s is coupled with the prediction that the population may increase to 7,000,000,000 by the year 2000, it becomes obvious that animal-food supplies must be increased. One way in which this might be accomplished is by learning to control the diseases that afflict animals throughout the world (see Table 11), especially in the developing nations of Asia and Africa, where the population is expanding most rapidly. Most of the information concerning animal diseases, however, applies to domesticated animals such as pigs, cattle, and sheep, which are relatively unimportant as food sources in these nations. Remarkably little is known of the diseases of the goat, the water buffalo, the camel, the elephant, the yak, the llama, or the alpaca; all are domesticated animals upon which the economies of many developing countries depend. It is in these countries that increased animal production resulting from the development of methods for the control and eradication of diseases affecting these animals is most urgently needed.

Despite the development of various effective methods of disease control, substantial quantities of meat and milk are lost each year throughout the world. In countries in which animal-disease control is not yet adequately developed, the loss of animal protein from disease is about 30 to 40 percent of the quantity available in certain underdeveloped areas. In addition, such countries also suffer losses resulting from poor husbandry practices.

Role in human disease. Animals have long been recognized as agents of human disease. Man has probably been bitten, stung, kicked, and gored by animals for as long as he has been on earth; in addition, early man sometimes became ill or died after eating the flesh of dead animals. In more recent times, man has discovered that many invertebrate animals are capable of transmitting causative agents of disease from man to man or from other vertebrates to man. Such animals, which act as hosts, agents, and carriers of disease, are important in causing and perpetuating human illness. Because about three-fourths of the important known zoonoses are associated with domesticated animals, including pets, the term zoonoses was originally defined as a group of diseases that man is able to acquire from domesticated animals. But this definition has been modified to include all human diseases (whether or not they manifest themselves in all hosts as apparent diseases) that are acquired from or transmitted to any other vertebrate animal. Thus, zoonoses are naturally occurring infections and infestations shared by man and other vertebrates. Although the role of domesticated animals in many zoonoses is understood, the role of the numerous species of wild animals with which man is less intimately associated is not well understood. The discovery that diseases such as yellow fever, viral brain infections, plague, and numerous other important diseases involving man or his domesticated animals are fundamentally diseases of wildlife and exist independently of man and his civilization, however, has increased the significance of studying the nature of wildlife diseases. Table 10 contains a partial list of zoonoses, including the causative agents and the animals involved.

Role of domesticated animals

Animals in research: the biomedical model. Although in modern times the practice of veterinary medicine has been separated from that of human medicine, the observations of the physician and the veterinarian continue to add to the common body of medical knowledge. Of the more than 1,200,000 species of animals thus far identified, only a few have been utilized in research, even though it is likely that, for every known human disease, an identical or similar disease exists in at least one other animal species. Veterinary medicine plays an ever-increasing role in the health of man through the use of animals as biomedical models with similar disease counterparts in man. This use of animals as models is important because research on many genetic and chronic diseases of man cannot be carried out using humans.

Importance of animals in research

Hundreds of thousands of mice and monkeys are utilized each year in research laboratories in the U.S. alone. Animal studies are used in the development of new surgical techniques (*e.g.,* organ transplantations), in the testing of new drugs for safety, and in nutritional research. Animals are especially valuable in research involving chronic degenerative diseases because they can be induced experimentally in them with relative ease. The importance of chronic degenerative diseases, such as cancer and cardiovascular diseases, has increased in parallel with the growing number of communicable diseases that have been brought under control. See Table 2 for a list of animals with diseases similar to those that occur in man.

Examples of animal diseases that are quite similar to commonly occurring human diseases include chronic emphysema in the horse; leukemia in cats and cattle; muscular dystrophies in chickens and mice; atherosclerosis in pigs and pigeons; blood-coagulation disorders and nephritis in dogs; gastric ulcers in swine; vascular aneurysms (permanent and abnormal blood-filled area of a blood vessel) in turkeys; diabetes mellitus in Chinese hamsters; milk allergy and gallstones in rabbits; hepatitis in dogs and horses; hydrocephalus (fluid in the head) and skin allergies in many species; epilepsy in dogs and gerbils; hereditary deafness in many small animals; cataracts in the eyes of dogs and mice; and urinary stones in dogs and cattle.

The study of animals with diseases similar to those that affect man has increased knowledge of the diseases in man; knowledge of nutrition, for example, based largely on the results of animal studies, has improved the health of animals, including man. Animal investigations have been used extensively in the treatment of shock, in open-heart surgery, in organ transplantations, and in the testing of new drugs. Other important contributions to human health undoubtedly will result from new research discoveries involving the study of animal diseases.

ROLE OF ECOLOGY

Epidemiology, the study of epidemics, is sometimes defined as the medical aspect of ecology, for it is the study of diseases in animal populations. Hence the epidemiologist is concerned with the interactions of organisms and their environments as related to the presence of disease. The multiple-causality concept of disease embraced by epidemiology involves combinations of environmental factors and host factors, in addition to the determination of the specific causative agent of a given disease. Environmental factors include geographical features, climate, and concentration of certain elements in soil and water. Host factors include age, breed, sex, and the physiological state of an animal as well as the general immunity of a herd resulting from previous contact with a disease. Epidemiology, therefore, is concerned with the determination of

Table 2: A Partial List of Biomedical Models in Veterinary Medicine

animal disease (model)	animal affected	human counterpart disease
Cardiovascular system diseases		
Hereditary lymphedema	dog	Milroy's disease
Elevated blood pressure	mouse	hypertension
Atherosclerosis	swine	atherosclerosis
Periarteritis nodosa	cattle	periarteritis nodosa
Dissecting aneurysms	turkey	aneurysm
High-altitude disease	cattle	right ventricular hypertrophy
Endocardial fibroelastosis	dog	endocardial fibroelastosis
Heart failure	dog	congestive heart failure
Congenital lymphatic edema	swine	lymphatic edema
Endocrine system diseases		
Diabetes mellitus	Chinese hamster	diabetes mellitus
Antidiuretic-hormone deficiency	mouse	diabetes insipidus
Polyuria	Chinese hamster	diabetes insipidus
Congenital goitre	cattle	goitre
Adrenal cortical hypertrophy	dog	hyperadrenocorticism
Snell's dwarf	mouse	thyrotropin deficiency
Adenohypophyseal aplasia	cattle	adenohypophyseal aplasia
Hyperinsulinism	dog	hyperinsulinism
Familial "adiposity"	mouse	obesity
Acetonemia	cattle	ketosis
Early senility	Syrian hamster	aging
Gastrointestinal system diseases		
Esophageal achalasia	dog	achalasia
Cleft palate	horse	cleft palate
Gastric ulcer	swine	gastric ulcer
Regional ileitis	swine	regional ileitis
Granulomatosis colitis	boxer dog	ulcerative colitis
Acute hemorrhagic colitis	rabbit	hemorrhagic colitis
Megacolon	mouse	megacolon
Pancreatitis	dog	pancreatitis
Liver diseases		
Viral hepatitis	subhuman primate	viral hepatitis
Serum hepatitis	horse	transfusion hepatitis
Dubin–Johnson syndrome	sheep	Dubin–Johnson syndrome
Congenital photosensitivity and hyperbilirubinemia	Southdown sheep	Gilbert's syndrome
Nonhemolytic hyperbilirubinemia	rat	Crigler–Najjar syndrome
Pigmentary liver disease	howler monkey	hepatocellular melanosis
Hepatorenal syndrome	dog	hepatorenal syndrome
Hepatic coma	horse	hepatic coma
Glycogen-storage syndrome	dog	von Gierke's syndrome
Lantana camara poisoning	sheep	kwashiorkor
Pyrrolizidine plant alkaloids	cattle	veno-occlusive disease
Hemopoietic system diseases		
Congenital erythrocytic porphyria (recessive)	cattle	congenital erythrocytic porphyria
Congenital porphyria (dominant)	cat	erythrocytic porphyria
Hereditary leukomelanopathy	mink	Chediak–Higashi syndrome
Pelger–Huët anomaly	cattle	Pelger–Huët anomaly
Cyclic neutropenia	dog	cyclic neutropenia
Aleutian disease	mink	multiple myeloma
Abnormal lipid in lymphoid tumours	mouse	Niemann–Pick disease
Viral leukemia	cat	lymphocytic leukemia
Multiple myeloma	dog	multiple myeloma
Bialbuminemia	swine	bialbuminemia
Hemophilia (factor VIII)	dog	hemophilia
Factor VII deficiency	dog	factor VII deficiency
Hemophilia-B-like disease	dog	Christmas disease
Hertwig's anemia	mouse	macrocytic anemia
Malaria	penguin	malaria
In vitro sickling of erythrocytes	deer	sickle-cell anemia
Muscle diseases		
Hereditary muscular dystrophy	chicken	muscular dystrophy
Muscle diseases (cont.)		
Polymyopathy	Syrian hamster	muscular dystrophy
Muscular dysgenesis	mouse	prenatal muscle degeneration
Nutritional muscular dystrophy	sheep	muscular dystrophy
Paralytic myoglobinuria	horse	paroxysmal myoglobinuria
Myoclonia congenita	swine	myotonia congenita
Nervous system diseases		
Cerebellar hypoplasia	cat	cerebellar hypoplasia
Nigropallidal encephalomalacia	horse	Parkinson's disease
Hydrocephalus	rabbit	hydrocephalus
Leukoencephalosis	mouse	dystrophy of white matter
Globoid leukodystrophy	dog	globoid leukodystrophy
Grand-mal seizures	gerbil	epilepsy
Lipodystrophy	dog	familial amaurotic idiocy
Scotty cramps	dog	neurogenic muscular cramps
Milk fever	cattle	hypocalcemia
Trembler mutation	mouse	tremours
Hereditary ataxia	calf	ataxia
Congenital myotonia	goat	myotonia
Eye and ear diseases		
Hereditary deafness	cat	deafness
Cochlear degeneration	mouse	cochlear degeneration
Hypoplasia of organ of Corti	dog	hypoplasia of organ of Corti
Hereditary glaucoma	rabbit	glaucoma
Inherited cataract	cattle	cataract
Hereditary iridal heterochromia	cattle	iridal heterochromia
Congenital retinal dysplasia	dog	retinal dysplasia
Retinal dystrophy	mouse	pigmented retina
Diabetic microaneurysms	dog	diabetic microaneurysms
Reproductive system diseases		
Toxemia of pregnancy	guinea pig	toxemia of pregnancy
Prolonged gestation	cattle	prolonged gestation
Uterine cystic hyperplasia	mouse	uterine cystic hyperplasia
Prostatic hyperplasia	canine	prostatitis
Cryptorchidism	swine	cryptorchidism
Respiratory system diseases		
Acute pulmonary emphysema	cattle	pulmonary emphysema
Chronic pulmonary emphysema	horse	pulmonary emphysema
Pulmonary adenomatosis	cattle	adenomatosis
Pneumonia	dog	Hechi's pneumonia
Induced lung tumours	mouse	lung tumours
Skeletal system diseases		
Osteodystrophy	primate	fibrous osteodystrophy
Familial osteoporosis	dog	osteogenesis imperfecti
Senile osteoporosis	mouse	senile osteoporosis
Achondroplasia	rabbit	dwarfism
Intervertebral-disk syndrome	dog	disk luxation
Hip dysplasia	dog	acetabular dysplasia
Clubfoot	mouse	clubfoot
Skin diseases		
Baldness, male pattern	stumptail macaque	baldness, male pattern
Albinism	mouse	albinism
Genetic hypotrichosis	cattle	hypotrichosis
Hyperkeratosis	cattle	hyperkeratosis
Cutis hyperelastica	dog	Ehlers–Danlos disease
Seborrheic dermatitis	dog	seborrheic dermatitis
Impetigo	dog	impetigo
Milia	dog	milia
Urinary system diseases		
Diabetes insipidus	mouse	diabetes insipidus
Cystinuria	blotched genet	cystinuria
Chronic interstitial nephritis	dog	uremia
Cystic or absent kidneys	rat	cystic kidneys
Renal amyloidosis	mouse	renal amyloidosis
Cloisonné kidneys	goat	renal hemosiderosis

the individual animals that are affected by a disease, the environmental circumstances under which it may occur, the causative agents, and the ways in which transmission occurs in nature. The epidemiologist, who utilizes many scientific disciplines (*e.g.*, medicine, zoology, mathematics, anthropology), attempts to determine the types of diseases that exist in a specific geographical area and to control them by modifying the environment.

Diseases in animal populations are characterized by certain features. Some outbreaks are termed sporadic diseases because they appear only occasionally in individuals within an animal population. Diseases normally present in an area are referred to as endemic, or enzootic, diseases, and they usually reflect a relatively stable relationship between the causative agent and the animals affected by it. Diseases that occasionally occur at higher than normal rates in animal populations are referred to as epidemic, or epizootic, diseases, and they generally represent an unstable relationship between the causative agent and affected animals.

Sporadic and endemic diseases

The effect of diseases on a stable ecological system, which is the result of the dominance of some plants and animals and the subordination or extinction of others, depends on the degree to which the causative agents of diseases and their hosts are part of the system. Epidemic diseases result from an ecological imbalance; endemic diseases often represent a balanced state. Ecological imbalance and, hence, epidemic disease may be either naturally caused or induced by man. A breakdown in sanitation in a city, for example, offers conditions favourable for an increase in the rodent population, with the possibility that diseases such as plague may be introduced into and spread among the human population. In this case, an epidemic would result as much from an alteration in the environment as from the presence of the causative agent *Pasteurella pestis,* since, in relatively balanced ecological systems, the causative agent exists enzootically in the rodents (*i.e.,* they serve as reservoirs for the disease) and seldom involves man. In a similar manner, an increase in the number of epidemics of viral encephalitis, a brain disease, in man has resulted from the ecological imbalance of mosquitoes and wild birds caused by man's exploitation of lowland for farming. Driven from their natural habitat of reeds and rushes, the wild birds, important natural hosts for the virus that causes the disease, are forced to feed near farms; mosquitoes transmit the virus from birds to cattle and man.

Detection and diagnosis

REACTIONS OF TISSUE TO DISEASE

As previously noted, disease may be defined as an injurious deviation from a normal physiological state of an organism sufficient to produce overt signs, or symptoms. The deviation may be either an obvious organic change in the tissue comprising an organ or a functional disturbance whose organic changes are not obvious. The severity of the changes that occur in cells and tissues subjected to injurious agents is dependent upon both the sensitivity of the tissue concerned and the nature and time course of the agent. A mildly injurious agent that is present for short periods of time may either have little effect or stimulate cells to increased activity. Strongly injurious agents in prolonged contact with cells cause characteristic changes in them by interfering with normal cell processes. Most causative agents of disease fall into the latter category. Causative agents and some of the symptoms of many of the diseases mentioned in this section are found in Tables 3 through 8.

Degeneration and infiltration

Characteristics of cell and tissue changes. Changes in cells and tissues as a result of disease include degenerative and infiltrative changes. Degenerative changes are characterized by the deterioration of cells or a tissue from a higher to a lower form, especially to a less functionally active form. When chemical changes occur in the tissue, the process is one of degeneration. When the changes involve the accumulation of materials within the cells comprising tissues, the process is called infiltration. Diseases such as pneumonia, metal poisoning, or septicemia (the persistence of disease-causing bacteria in the bloodstream) may cause the mildest type of degeneration—parenchymatous changes, or cloudy swelling of the cells; the cells first affected are the specialized cells of the liver and the kidney. Serious cellular damage may cause the uptake of water by cells (hydropic degeneration), which lose their structural features as they fill with water. The causes for the accumulation in cells of abnormal amounts of fats (fatty infiltration and degeneration) have not yet been established with certainty but probably involve fat metabolism. Poisons such as phosphorus may cause sudden increases in the accumulation of fats in the liver. An abnormal protein material may accumulate in connective-tissue components of small arteries as a result of chronic pneumonia, chronic bacterial infections, and prolonged antitoxin production (in horses); the condition is known as amyloid degeneration and infiltration. Hyaline degeneration, characterized by tissues that become clear and appear glasslike, usually occurs in connective-tissue components of small blood vessels as a result of conditions that may occur in kidney structures (glomeruli) of animals with nephritis or in lymph glands of animals with tuberculosis. Certain structures (glomeruli) of animals with nephritis result in degeneration.

The condition in which mucus, a secretion of mucous membranes lining the inside surfaces of organs, is produced in excess and accumulates in greater than normal amounts is referred to as mucoid degeneration. Major causes of this condition include chronic irritation of mucous membranes and certain mucus-producing tumours. Abnormal amounts of glycogen, which is the principal storage carbohydrate of animals, may occur in the liver as a result of certain inherited diseases of animals; the condition is known as glycogen infiltration. The abnormal deposition of calcium salts, which is known as hypercalcification, may occur as a result of several diseases involving the blood vessels and the heart, the urinary system, the gallbladder, and the bonelike tissue called cartilage. Pigments (coloured molecules) from coal dust or asbestos dust may infiltrate the lungs of certain dogs in two types of lung disease: anthracosis and asbestosis. Abnormal amounts of iron-containing coloured molecules (hemosiderin) resulting from the breakdown of hemoglobin, the oxygen-carrying protein of red blood cells, are often deposited in the liver and the spleen after diseases that involve excessive breakdown of red blood cells. A dark-coloured molecule (melanin) occurs abnormally in the livers of certain sheep suffering from Dubin–Johnson syndrome and in certain tumours called melanomas. Uric acid infiltration, which occurs in poultry, is characterized by the deposition of uric acid salts.

Necrosis, the death of cells or tissues, takes place if the blood supply to tissues is restricted; poisons produced by microbes, chemical poisons, and extreme heat or electricity also may cause necrosis. The rotting of the dead tissue is known as gangrene.

Atrophy of animal tissue involves a process of tissue wasting, in which a decrease occurs in the size or number of functional cells—*e.g.,* in inherited muscular dystrophy of chickens. Hypertrophy—an increase in the size of the cells in a tissue or an organ—occurs in heart muscle during diseases involving the heart valves, in certain pneumonias, and in some diseases of the endocrine glands. Aplasia is the term used when an entire organ is missing from an animal; hypoplasia indicates arrested or incomplete development of an organ, and hyperplasia an increase in the production of the number of cells—*e.g.,* the persistent callus that forms on the elbows of some dogs. Metaplasia is used to describe the change of one cell type into another; it may occur in chronic irritation of tissues and in certain cancerous tumours.

Characteristics of inflammatory reactions. When tissues are injured, they become inflamed. The inflammation may be acute, in which case the inflammatory processes are active, or chronic, in which case the processes occur slowly and new connective tissue is formed. The reaction of inflamed tissues is a combination of defensive and repair mechanisms. Acute inflammation is characterized by redness, heat, swelling, sensitivity, and impaired function. Several types of acute inflammation are known. Mild acute inflammations of mucous membranes resulting in the production of thin watery material (exudate) are called catarrhal inflammations; parenchymatous inflammations occur in organs undergoing degeneration. If the exudate formed in response to an injury is of a serous nature—that is, resembling blood plasma—the process is called serous inflammation. In fibrinous inflammation, a protein (fibrin) forms on membranes, including those in the lungs. In suppurative inflammation, dead tissue is replaced with pus composed of colourless blood cells (leucocytes) and tissue juices.

During the inflammatory reaction, the injured tissue is surrounded by an area of rapidly dividing cells. Specialized cells called macrophages enter the tissue and remove blood and tissue debris. Other cells, called neutrophils, ingest disease-causing bacteria and other foreign material. In chronic inflammations, the connective tissue contains fibroblasts, cells that divide and form new connective, or scar, tissue.

Table 3: Selected Infectious and Parasitic Diseases of Animals

animal(s) affected	name(s) of disease	causative organism	nature of disease
Diseases of bacterial origin			
Most mammals, chickens	necrobacillosis, calf diptheria, bovine foot rot, necrotic hepatitis, dermatitis	*Sphaerophorus necrophorus*	organism invades tissue and causes tissue death (necrosis) after other wounds or infections have occurred; *i.e.*, disease is known as a secondary infection
Cattle, sheep, horses, chickens, man, many other animals	botulism	toxins produced by *Clostridium botulinum*	results from eating toxins released in decayed or spoiled foods; toxins cause rapid paralysis of nerves in throat and all muscles, almost always fatal
Swine, cattle, sheep, goats, rabbits, man, dogs, many other animals	listeriosis, circling disease, meningoencephalitis	*Listeria monocytogenes*	symptoms vary in affected animal; organism may affect the central nervous system (brain, spinal cord) or the membranes surrounding it or cause necrosis of heart muscles, localized tissue death in liver, or a septicemia (persistence of bacteria in the bloodstream)
Swine, cattle, sheep, goats, horses, turkeys, man	erysipelas, diamond-skin disease	*Erysipelothrix insidiosa*	manifestations include septicemia and pathological discontinuities of tissue (lesions) in skin, heart, joints (in swine); arthritis (in sheep, occasionally in cattle, horses, goats); septicemia and death (in turkeys); skin lesions (in man)
Cattle, sheep, goats, horses, mules	anthrax, splenic fever, charbon	*Bacillus anthracis*	spores (inactive forms) of organisms in soil, transmitted through insect bites or food; manifestations include hemorrhage and edema (accumulation of fluid) in tissues
Swine, cattle, sheep, horses, mules	malignant edema, gas gangrene	*Clostridium septicum*	spores enter from dirt into injured tissue, cause severe gangrene (rotting of dead tissue), swelling; prognosis (outlook) poor
Swine, cattle, sheep, goats	pyobacillosis	*Corynebacterium pyogenes*	characterized by multiple abscesses (localized collections of pus) throughout the body; may result in debilitation (including arthritis in swine) and death
Primarily swine, cattle, goats (secondarily in man and other animals)	brucellosis, Bang's disease, contagious abortion; undulant fever, Malta fever (in man)	*Brucella abortus, B. melitensis, B. suis.*	primarily affects genital organs in both sexes; may cause abortion, sterility, infection of fetus in female, local lesions in various tissues; pasteurization of milk has controlled the disease in man
Swine, cattle, sheep	shipping fever, pasteurellosis, hemorrhagic septicemia	*Pasteurella multocida* and *P. hemolytica;* also in conjunction with viral agents	causes of great economic losses throughout the world; manifestations may include acute to chronic respiratory disease; the various causative organisms vary in virulence (degree of pathogenicity); an acute form (*i.e.*, short, severe) affects rabbits
Horses, mules, donkeys (man less susceptible)	glanders, farcy, malleus	*Malleomyces mallei*	organisms enter animal through digestive tract, travel via blood to lungs, trachea, and skin, and form ulcers; an acute form causes death, a chronic type may persist for years; human infections occur from exposure of broken skin to affected animals
Horses (mules, donkeys less susceptible)	strangles, distemper, infectious adenitis	*Streptococcus equi*	most common in young, undernourished horses in crowded conditions; manifested by high temperature, nose infections, and abscesses in lymph glands of neck
Horses	purpura hemorrhagica, petechial fever	unknown, but associated with *Streptococcus equi*	noncontagious, follows acute infections and toxemias; characterized by generalized hemorrhages in tissues and the accumulation of fluid (edema); relapses often occur
Primarily horses	ulcerative lymphangitis or cellulitis	*Corynebacterium pseudotuberculosis*	chronic disease, develops slowly following the entrance of bacteria through skin; affects hindlegs, sometimes severely
Horses (males most susceptible); sometimes swine, cattle; rarely dogs or cats	tetanus, lockjaw	toxins produced by *Clostridium tetani*	bacteria enter tissue at time of injury, produce toxins in necrotic tissue; affected animals become stiff; death results from suffocation
Swine	streptococcal infection	*Streptococcus* species	younger pigs more easily infected; symptoms varied (*e.g.*, septicemia, arthritis, uterine inflammation, middle-ear infection, multiple abscesses)
Swine, accidentally in other animals	salmonellosis, enteritis, swine typhoid	*Salmonella choleraesuis*	may be acute (*i.e.*, have a short, severe course) or chronic (persists for a long time); symptoms include loss of weight, sometimes acute septicemia
Swine, cattle	actinobacillosis, botryomycosis, big head	*Actinobacillus lignieresii*	organism a normal inhabitant of mouth, enters tissues through ulcers or wounds; symptoms include abscesses
Cattle	leptospirosis, hemoglobinuria	*Leptospira* species	symptoms of acute form include abortion, bloody milk, hemoglobin (blood pigment) in urine, kidney disease, and destruction of red blood cells; a milder form also exists
Cattle	infectious bovine pyelonephritis, infectious cystitis	*Corynebacterium renale*	usually observed in pregnant cattle in winter; a slowly developing disease that affects kidneys and bladder
Primarily cattle (rarely man, swine, sheep, horses)	tuberculosis, pearly disease	*Mycobacterium tuberculosis*	a chronic disease characterized by lesions, usually in lungs and lymph nodes, but sometimes in many other organs
Cattle (rarely sheep)	bacillary hemoglobinuria, red water disease	*Clostridium hemolyticum*	spores eaten with food, develop into active cells, migrate to liver, and produce infarcts (tissue death); usually fatal within 36 hours

Table 3: Selected Infectious and Parasitic Diseases of Animals (continued)

animal(s) affected	name(s) of disease	causative organism	nature of disease
Cattle, sheep	pinkeye, infectious keratitis, kerato-conjunctivitis	*Moraxella bovis* (in cattle), *Colesiota conjunctivae* (in sheep)	affects eyes; may result in blindness; a very contagious disease whose spread may be influenced by dust irritation or possibly by viral invasion
Cattle, sheep	Johne's disease, paratuberculosis	*Mycobacterium paratuberculosis*	a chronic disease; causes diarrhea, progressive weight loss
Cattle, sheep	vibriosis, epizootic abortion	*Vibrio fetus*	a venereal disease (*i.e.*, transmitted by sexual contact) in cattle; transmitted in contaminated food and water in sheep; commonly results in infertility or abortion in cattle, abortion in sheep
Cattle, sheep (occasionally swine, goats, deer, horses)	blackleg, black quarter, quarter ill	*Clostridium feseri* (*chauvoei*)	spores transmitted from soil to animal through wounds or cuts; symptoms include lameness, gangrene of affected tissues (usually in leg muscles); usually fatal
Sheep	enterotoxemia, over-eating disease, pulpy kidney	*Clostridium perfringens* type D	affected lambs usually fat or feeding on rich clover pasture; usually fatal within a day from acute toxemia (absorption of bacterial toxins)
Primarily sheep	pseudotuberculosis, caseous lymphadenitis	*Corynebacterium ovis*	organisms transmitted to animal through breaks in the skin; slowly developing abscesses (usually in lungs or lymph nodes) may rupture and spread throughout body
Sheep, goats (occasionally cattle)	black disease, infectious necrotic hepatitis	*Clostridium novyi*	organisms probably present normally in intestinal tract, associated with fluke (parasitic worm) movements in liver; produce liver damage, toxemia, and death
Diseases of viral origin			
Mammals	rabies, hydrophobia, lyssa, mad dog, le Rage	rabies virus	transmitted primarily through the bite of a rabid animal; wild animals (*e.g.*, skunks, squirrels, bats) a reservoir for infection; disease characterized by central-nervous-system symptoms (*e.g.*, rage, excitability; paralysis of jaw with salivation), general paralysis, and death
Mammals (especially cattle)	bovine warts, papillomatosis	papilloma viruses	warts of variable size develop, usually on sides of head and neck of cattle, sometimes on sex organs
Many mammals (*e.g.*, swine, cattle, sheep, goats, horses)	pox, variola	pox virus	often an acute highly infectious disease, characterized by formation of papules (small solid elevations), vesicles (small liquid-containing sacs), and pustules (small pus-filled elevations) on the skin
Swine, cattle (also rats, dogs, cats)	pseudorabies, Aujeszky's disease, mad itch	pseudorabies virus	affected animals rub body parts, undergo spasmodic muscle contractions; froth at the mouth; and show nervous irritability; usually fatal
Young pigs	transmissible gastroenteritis (TGE)	TGE virus	acute and fatal to pigs less than two to three weeks old; virus attacks absorptive surfaces of small intestine
Swine	hog cholera, swine fever, swine pest	hog-cholera virus	infectious disease; may be acute or chronic; spread by flies, animal contact, garbage, contaminated pastures; symptoms include high fever, severe hemorrhages in skin and organs
Swine	vesicular exanthema (VE)	VE virus	vesicles form on snout, mouth, abdominal wall; foot lesions occur; highly infectious, spread through animal contact or raw pork scraps in garbage; fatal usually only in young pigs
Swine, ferrets, mice	swine influenza, hog flu	swine-influenza virus; bacterium: *Hemophilus suis*	acute contagious disease; virus enters animal through lungworm larvae, acute infection occurs if *Hemophilus* organisms are in lung; symptoms include fever, pneumonia, bronchitis
Swine, cattle, horses	vesicular stomatitis (VS), mouth thrush	VS virus	effects varied; *e.g.*, high fever, salivation, vesicles in mouth region, lack of appetite (in horses); inflammation of mammary glands, vesicles in mouth region, inflammation of feet (in cattle); snout and mouth lesions, severe feet involvement (in swine)
Cattle	sporadic bovine encephalomyelitis (SBE), Buss disease	SBE virus	weakens calves; an infection of brain and membranes of brain and spinal cord; recovery often occurs
Cattle	infectious bovine rhinotracheitis (IBR), red nose, pinkeye, dust pneumonia	IBR virus	acute infection followed by secondary bacterial infections (*Pasteurella multocida*, *Spherophorus necrophorus*) in lungs, sex organs, eye; nostrils swell, become red
Cattle, buffalo, deer	malignant catarrhal fever (MCF), head catarrh, snotsiekte, epitheliosis	MCF virus	numerous symptoms include rapid weight loss, eye lesions, nasal discharge, muscular twitching, convulsions; usually fatal
Cattle, sheep	bluetongue, soremuzzle, catarrhal fever	bluetongue virus	virus transmitted through gnats (small flies); a serious problem in sheep; numerous symptoms; recovery very slow; usually less than 10 percent mortality of a flock
Sheep (also transmissible to cows)	ovine virus abortion (OVA), enzootic abortion	OVA virus	causes economic losses through abortion, weak lambs, and poor breeding efficiency

Table 3: Selected Infectious and Parasitic Diseases of Animals (continued)

animal(s) affected	name(s) of disease	causative organism	nature of disease
Sheep, goats, man	contagious ecthyma (CE), sore mouth, doby mouth, orf, pustular dermatitis	CE virus	udder, lips, and nose of sheep affected; secondary bacterial invasion may result in death, but animals usually recover
Horses	equine infectious anemia (EIA), swamp fever, malarial fever	EIA virus	transmitted by mosquitoes, lice, flies, and hypodermic needles; either acute, chronic, or latent (not manifest); varied symptoms include intermittent fever, loss of weight, jaundice, hemorrhages, anemia, and fluid in body cavities
Horses	equine viral arteritis (EVA), infectious arteritis	EVA virus	an acute contagious disease similar to EVR in symptomatology but causes damage to small arteries
Horses, mules, man, laboratory animals	equine encephalomyelitis (EE), sleeping sickness, viral encephalitis	EE virus	many viral strains transmitted by an insect (usually mosquitoes or mites or ticks); disease causes inflammation of brain cells; the many symptoms include death; inapparent infections occur in chickens, pigeons, and pheasants
Horses (also guinea pigs, mice, hamsters)	equine viral rhinopneumonitis (EVR), equine virus abortion	EVR virus	disease highly contagious; symptoms include high fever, mild upper-respiratory involvement, and usually abortion with liver damage of fetus in pregnant mares
Diseases of fungal origin			
Many domestic, laboratory, and wild mammals	histoplasmosis, reticuloendothelial cytomycosis	*Histoplasma capsulatum*	chronic; may resemble tuberculosis; granulomas (tumours) in lungs, liver, and spleen; intestinal involvement in dogs results in diarrhea
Swine, cattle, sheep, horses, fowl, dogs, cats	ringworm, tinea, trichophytosis	*Trichophyton* and *Microsporum* species	infectious skin disease caused by invasion of hair follicles; characterized by round crusty lesions, inflammation
Swine, cattle, horses (man secondarily)	actinomycosis, lumpy jaw, wooden tongue	*Actinomyces bovis*	cattle manifest a bonelike swelling on upper or lower jaw; swine manifest a tumourlike enlargement of udder caused by infections from teeth of suckling pigs
Cattle, horses, cats, dogs, man	cryptococcosis	*Cryptococcus neoformans*	usually caused by inhalation of contaminated dust; lungs affected primarily; disease may spread to almost any organ
Cattle, sheep, dogs, man	coccidioidomycosis coccidioidal granuloma, San Joaquin Valley fever	*Coccidioides immitis*	in the acute respiratory form, symptoms include cough, with recovery in two weeks; in the more serious chronic form, gradual loss in weight, abscesses, and granulomas in various tissues, including skin, occur
Primarily horses and man	sporotrichosis	*Sporotrichum schenckii*	occurs first as ulcers on skin, invasion of the lymph glands occurs, may spread throughout circulatory system
Diseases of rickettsial origin			
Swine	eperythrozoonosis, ictero-anemia, yellow-belly	*Eperythrozoon suis*	organisms cause red-blood-cell destruction; both acute and mild forms; many animals with a mild form act as carriers
Cattle	anaplasmosis, South African gall sickness	*Anaplasma marginale*	infectious disease spread either by blood-sucking ticks, mosquitoes, flies, or by mechanical transmission (*e.g.*, resulting from dehorning, vaccination); symptoms usually include extreme anemia from destruction of red blood cells; recovered animals are immunological carriers
Diseases of protozoal origin			
Most animals (including man)	toxoplasmosis	*Toxoplasma gondii*	transmission not clear but probably occurs by contaminated food or direct contact; symptoms include weakness, respiratory problems, lack of coordination, nodules throughout body, enlarged lymph nodes, and tissue death; treatment difficult
Swine, cattle, sheep, goats	coccidiosis	*Eimeria zürnii, E. bovis*, and ten other species	causative organisms found in most mature animals; symptoms result in loss of large amounts of blood and dehydration; mortality may be as high as 50 percent
Cattle	trichomoniasis	*Trichomonas fetus*	transmitted by sexual contact or by artificial insemination, symptoms include abortion, failure to conceive, inflammation of uterus; no symptoms apparent in infected bull that acts as a carrier
Cattle	Texas fever, cattle-tick fever, babesiasis, piroplasmosis, red water	*Babesia bigemina*	organism, which destroys red blood cells, is transmitted by ticks (*Margaropus* species) and mechanical means (*i.e.*, surgical instruments, needles); symptoms include high fever, severe anemia, hemoglobin in urine; chronic forms occur; some animals act as carriers
Horses	dourine, equine syphilis, breeding disease	*Trypanosoma equiperdum*	transmitted by sexual contact or blood-sucking flies; affects sex organs, causes plaquelike areas on skin, paralysis of muscles, loss of condition
Horses, mules, donkeys	equine piroplasmosis, equine malaria, babesiasis	*Babesia caballi, B. equi*	acute cases may die quickly; animals with less acute forms have varied symptoms (*e.g.*, intermittent fever, jaundice, internal hemorrhages); anemia results from invasion of red blood cells by causative organisms; *B. equi* more pathogenic than *B. caballi*

Table 3: Selected Infectious and Parasitic Diseases of Animals (continued)

animal(s) affected	name(s) of disease	causative organism	nature of disease
Diseases of nematode (roundworm) origin			
Swine	lungworms	*Metastrongylus* species	common symptoms include coughing and lung irritation
Swine	kidney worms	*Stephanurus dentatus*	mature worms live in urinary tract; larvae migrate to liver, produce lesions and weight loss
Swine	intestinal roundworm	*Ascaris suum*	migrations of organisms through lungs cause hemorrhages and pneumonia, may interfere with bile flow and food absorption
Swine	intestinal threadworm	*Strongyloides*	migration of large numbers of larvae cause tissue damage
Cattle	stomach worms	*Ostertagia* and *Trichostrongylus* species	symptoms include anemia, stunted growth, and diarrhea
Cattle	nodular worm	*Oesophagostomum radiatum*	nodules form in tissues; poor intestinal absorption caused by larvae (immature forms of organism) and nodules results in diarrhea
Cattle	verminous pneumonia	*Dictyocaulus viviparus*	ingested infective larvae migrate to lungs, produce coughing, discomfort, and pneumonia
Sheep	lungworms	*Dictyocaulus filaria, Muellerius capillaris*	symptoms include formation of lung nodules, collapse of portions of lungs
Sheep	hookworms	*Bunostomum trigonocephalum*	bloodsucking hookworms cause anemia, intermittent diarrhea
Sheep	filarial dermatitis	*Elaeophora schneideri*	symptoms include skin lesions
Horses	oxyuriasis (pinworms)	*Oxyuris equi*	worms cause irritation in the area around the anus
Horses	ascariasis (intestinal roundworms)	*Parascaris equorum*	larval forms cause damage; symptoms include defective intestinal absorption
Horses	strongylosis	*Strongylus* species	large numbers of worms weaken foals (newborn horses); migrating larvae may cause formation of clots in blood vessels and lameness
Horses	habronemiasis (summer sores)	*Habronema* species	habronema larvae may enter skin wounds, causing granulation; eye and stomach inflammations also occur
Diseases of platyhelminth (flatworm) origin			
Sheep	tapeworm	*Moniezia expansa*	results in poor growth
Sheep	fringed tapeworm	*Thysanosoma actinoides*	causes digestive disturbances
Horses	tapeworm	*Anaplocephala perfoliata*	symptoms may include inflammation of gut and ulceration
Diseases of acanthocephalan (spiny-headed-worm) origin			
Swine	spiny-headed worm	*Macracanthorhynchus hirudinaceus*	nodules form on small intestine; may result in peritonitis (inflammation of lining of internal organs)
Diseases of arthropod origin			
Cattle	scabies (mange)	*Chorioptes bovis, Psoroptes equi, Sarcoptes scabiei* (mites)	contagious skin diseases
Cattle	grubs	*Hypoderma bovis* (heel fly)	migrating larvae produce tissue damage, cysts, and hide damage
Cattle	lice	*Linognathus vituli, Solenopotes capillatus, Haematopinus quadripertusis* (bloodsucking lice), *Bovicola bovis* (biting louse)	symptoms include dermatitis and anemia
Sheep	screwworm infestations	many fly larvae (*e.g., Cochliomyia hominivorax, Chrysomyia bezziana*)	flies lay eggs in open wounds; developing larvae (screwworms) burrow into tissue and destroy it
Sheep	psoroptic mange (sheep scab)	*Psoroptes communis* (mite)	all parts of skin inflamed, particularly those covered with wool

Hemorrhage and edema

Characteristics of circulatory disturbances. An increase in the rate of blood flow to a body part, which is referred to by the term congestion, or hyperemia, occurs during inflammation; a diminished blood flow to tissues is referred to by the term ischemia, or a local anemia. Examples of hemorrhage, the escape of blood from vessels, include epistaxis, or nosebleeds, in racehorses; hematemesis, or regurgitation of blood, in dogs with uremia; hemoptysis, or blood loss from lungs; hematuria, or blood in urine, of cattle with inflammation of the urinary bladder. Edema, a condition that is characterized by abnormal accumulations of fluid in tissues, occurs not only in a tissue during inflammation but also over the entire body if the concentration of blood-serum proteins, especially albumin, is low. A thrombosis, which is a blood clot in a blood vessel, may block or slow circulation of blood to tissues; if blood vessels become blocked, the condition is known as an embolism. The term infarction describes the necrosis that occurs in tissues whose blood supply is blocked by an embolism.

METHODS OF EXAMINATION

Before an unhealthy animal receives treatment, an attempt is made to diagnose the disease. Both clinical findings, which include symptoms that are obvious to a nonspecialist and clinical signs that can be appreciated only by a veterinarian, and laboratory test results may be necessary to establish the cause of a disease. A clinical examination should indicate if the animal is in good physical condition, is eating adequately, is bright and alert, and is functioning in an apparently normal manner. Many disease processes are either inflammatory or result from tumours. Malignant tumours (*e.g.,* melanomas in horses, squamous cell carcinomas in small animals) tend to spread rapidly and usually cause death. Other diseases cause the circulatory disturbances or the degenerative and infiltrative changes that are summarized in the preceding section. If a specific diagnosis is not possible, the symptoms of the animal are treated.

A case record of the information pertaining to an animal (or to a herd of animals) that is suspected of having a

Table 4: Examples of Noninfectious Diseases of Animals

animal(s) affected	name(s) of disease	nature of disease
Hereditary diseases		
Pigs, calves, foals	congenital absence of skin (epitheliogenesis imperfecti)	complete absence of skin over parts of body; fatal a few days after birth
Swine, cattle	congenital porphyria (pink tooth)	causes anemia, wine-coloured urine; results from a biochemical defect in the metabolism of a component (porphyrin) of the iron-containing pigment (hematin) of hemoglobin, the oxygen-carrying protein of blood
Cattle	prolonged gestation (prolonged pregnancy)	may cause a three-week to three-month delay in birth of calves, which when born are either large or deformed; a special type in Guernsey and Jersey breeds results in death of calves at birth
Metabolic diseases		
Cattle (rarely sheep and pigs)	milk fever (parturient paresis)	caused in lactating cattle by loss of calcium into the milk; low levels of calcium in blood cause muscular weakness, circulatory collapse, and loss of consciousness; treatment includes replacing calcium
Cattle, sheep	ketosis (acetonemia in cattle; pregnancy toxemia in sheep)	occurs in lactating cattle following calving and in sheep in terminal stage of pregnancy (both times of increased need for carbohydrates); causes paralysis and death; complex treatment includes replacing carbohydrates
Horses	azoturia (paralytic myoglobinuria)	paralytic disease of unknown cause; occurs during exercise following a period of inactivity; muscles degenerate; results in paralysis, dark-red urine
Functional diseases		
Cattle, sheep	kidney and bladder calculi (urolithiasis)	cattle eating range grasses with high silicon content may develop obstructions (solid masses containing silicon and protein) in urinary system; similar effects may occur with diets high in phosphate, in which case the solid mass contains magnesium ammonium phosphate and protein
Cattle, sheep	bloat (ruminal tympany)	distension (caused by gases) of first two stomachs of cows; conditions preventing eructation (belching) of gases are major causes; occurs primarily when cattle overeat on leguminous pastures (alfalfa and clovers); often fatal
Cattle, horses	pulmonary emphysema (heaves in horses)	acute form occurs in cattle, chronic form in horses; alveoli (small terminal air sacs) in lung rupture, reducing surface area for oxygen transport; causes not yet clear, but disease may follow pneumonia and allergic reactions
Nutritional deficiency diseases		
Most animals	vitamin A deficiency	caused by dietary insufficiency of vitamin A or substance from which vitamin A is formed (carotene); numerous manifestations in young and adult animals include night blindness
Pigs	iron deficiency	caused by insufficient iron in diet; manifestations include severe anemia and poor growth
Diseases caused by chemical agents		
Pigs, cattle, horses	bracken-fern poisoning	fern contains thiaminase, which destroys vitamin B_1 (thiamine), thereby producing a vitamin deficiency in horses and swine; in cattle the bone marrow is affected, and deficiency of blood cells and excessive hemorrhages occur
Cattle, sheep, horses	rye-grass staggers	manifestations include either liver degeneration and photosensitization or uncoordination, convulsions, and paralysis; cause not yet established, but fungus on the rye grass plays some role in the liver degeneration
Cattle (occasionally other animals)	sweet-clover poisoning	caused by eating moldy sweet-clover hay, which contains the anticoagulant compound dicoumarol; manifestations include extensive hemorrhages and severe blood loss after injury
Cattle, sheep	molybdenum poisoning	results if pasture soil contains toxic quantities (three parts per million) of molybdenum, which replaces copper in body; symptoms include diarrhea, loss of hair colour, anemia
Diseases caused by physical agents		
Cattle	hardware disease (traumatic reticuloperitonitis)	objects such as nails and small pieces of balling wire may be eaten with feed; perforation and inflammation of first stomach (reticulum) may occur; surgery sometimes necessary
Cattle	brisket disease (mountain sickness)	occurs in cattle at high altitudes where levels of atmospheric oxygen are too low to provide oxygen required; manifestations include enlargement of right heart, congestive heart failure

disease is begun at the time the animal is taken to a veterinarian (or the veterinarian visits the animal) and is continued through treatment. It includes a description of the animal (age, species, sex, breed); the owner's report; the animal's history; a description of the preliminary examination; clinical findings resulting from an examination of body systems; results of specific laboratory tests; diagnosis regarding a specific cause for the disease (etiology); outlook (prognosis); treatment; case progress; termination; autopsy, if performed; and the utilization of scientific references, if applicable.

The veterinarian must diagnose a disease on the basis of a variety of examinations and tests, since he obviously cannot interrogate the animal. Methods used in the preparation of a diagnosis include inspection—a visual examination of the animal; palpation—the application of firm pressure with the fingers to tissues to determine characteristics such as abnormal shapes and possible tumours, the presence of pain, and tissue consistency; percussion—the application of a short, sharp blow to a tissue to provoke an audible response from body parts directly beneath; auscultation—the act of listening to sounds that are produced by the body during the performance of functions (*e.g.*, breathing, intestinal movements); smells—the recognition of characteristic odours associated with certain diseases; and miscellaneous diagnostic procedures, such as eye examinations, the collection of urine, and heart, esophageal, and stomach studies.

General inspection. Deviation of various characteristics from the normal, observation of which constitutes the general inspection of an animal, is a useful aid in diagnosing disease. The general inspection includes examination of appearance; behaviour; body condition; respiratory movements; state of skin, coat, and abdomen; and various common actions.

The animal's appearance

The appearance of an animal may be of diagnostic significance; small size in a pig may result from retardation of growth, which is caused by hog-cholera virus (Table 3). Observation of the behaviour of an animal is of value in diagnosing neurological diseases; *e.g.*, muscle spasms occur in lockjaw (tetanus) in dogs, nervousness and convulsions in dogs with distemper (Table 5), dullness in horses with equine viral encephalitis (Table 3), and excitement in animals suffering from lead poisoning. Subtle behavioral changes may not be noticeable. The general condition of the body is of value in diagnosing diseases that cause excessive leanness (emaciation), including certain cancers, or other chronic diseases, such as a deficiency in the output

of the adrenal glands or tuberculosis. Defective teeth also may point to malnutrition and result in emaciation.

The respiratory movements of an animal are important diagnostic criteria; breathing is rapid in young animals, in small animals, and in animals whose body temperature is higher than normal. Specific respiratory movements are characteristic of certain diseases—*e.g.*, certain movements in horses with heaves (emphysema) or the abdominal breathing of animals suffering from painful lung diseases. The appearance of the skin and hair may indicate dehydration by lack of pliability and lustre; or the presence of parasites such as lice, mites, or fleas; or the presence of ringworm infections and allergic reactions by the skin changes they cause. The poisoning of sheep by molybdenum in their hay may be diagnosed by the loss of colour in the wool of black sheep. Distension of the abdomen may indicate bloat in cattle or colic in horses.

Abnormal activities may have special diagnostic meaning to the veterinarian. Straining during urination is associated with bladder stones; increased frequency of urination is associated with kidney disease (nephritis), bladder infections, and a disease of the pituitary gland (diabetes insipidus). Excessive salivation and grinding of teeth may be caused by an abnormality in the mouth. Coughing is associated with pneumonia. Some diseases cause postural changes: for example, a horse with tetanus may stand in a stiff manner. An abnormal gait in an animal made to move may furnish evidence as to the cause of a disease, as louping ill in sheep.

Clinical examination. Following the general inspection of an animal thought to have contracted a disease, a more thorough clinical examination is necessary, during which various features of the animal are studied. These include the visible mucous membranes (conjunctiva of the eye, nasal mucosa, inside surface of the mouth, and tongue); the eye itself; and such body surfaces as the ears, horns (if present), and limbs. In addition, the pulse rate and the temperature are measured.

Structural features involved

The veterinarian examines the visible mucous membranes of the eye, nose, and mouth to determine if jaundice, hemorrhages, or anemia are present. The conjunctiva, or lining of the eye, may exhibit pus in pinkeye infections, have a yellow appearance in jaundice, or exhibit small hemorrhages in certain systemic diseases. Examination of the nose may reveal ulcers and vesicles (small sacs containing liquid), as in foot-and-mouth disease, a viral disease of cattle, or vesicular exanthema, a viral disease of swine. Ulceration of the tongue may be apparent in animals suffering from actinobacillosis, a disease of bacterial origin (see Table 3).

A detailed examination of the eye may show abnormalities of the cornea resulting from such diseases as infectious hepatitis in dogs (Table 5), bovine catarrhal fever, and equine influenza. Cataract, a condition in which the passage of light through the lens of the eye is obstructed, may result from a disorder of carbohydrate metabolism (diabetes mellitus), infections, or a hereditary defect.

An elevated temperature, or fever, resulting from the multiplication of disease-causing organisms may be the earliest sign of disease. The increase in temperature activates the body mechanisms that are necessary to fight off foreign substances. Measuring the pulse rate is useful in determining the character of the heartbeat and of the circulatory system.

TESTS AS DIAGNOSTIC AIDS

In many cases, the final diagnosis of an animal disease is dependent upon a laboratory test. Some involve mea-

Table 5: Some Common Diseases of Dogs

name(s) of disease	causative agent	nature of disease
Distemper	virus	affects nonvaccinated (nonimmunized) puppies in contact with infected animals; symptoms include loss of appetite, fever; inflammation of the brain is usual cause of death; some dogs may recover, but others have spastic tremors; foxes, wolves, mink, skunks, raccoons, and ferrets also susceptible
Infectious hepatitis (Rubarth's disease)	virus	affects dogs by causing hemorrhages and severe liver damage; affects foxes by causing inflammation of the brain; clinical signs are variable because disease symptoms vary from severe to inapparent (*i.e.*, no manifest signs)
Salmon poisoning	rickettsia	occurs after consumption of raw salmon or trout carrying rickettsial-infected flatworm (fluke) larvae (*Nanophyetus salmincola*); affects dogs, foxes, and coyotes primarily in the Pacific northwestern United States; symptoms include high fever, swollen lymph nodes; usually fatal within five days
Prostatitis	varied	inflammation of a gland near the urinary bladder (prostate gland) in male dogs; usually controlled by antibiotic drugs; other prostate-gland disorders may result from tumours (carcinoma, sarcoma) or from abnormal increase in cell multiplication (hyperplasia)
Congenital heart disease	inherited tendency	may occur in 1 percent of all dogs; heart disorders may lead to secondary diseases such as pneumonia, accumulation of fluid in body cavities, laboured breathing, edema; heart failure occurs
Hip dysplasia	apparently inherited tendency	crippling disorder common in many breeds (especially German shepherds); a shallow hip socket (acetabulum) results in an unstable hip joint, particularly during motion of hindleg
Kidney stones (calculi, urolithiasis)	hereditary, functional disturbance	calculi develop in kidney, bladder, and male urethra (tube from bladder to outside of body); surgery usually necessary; inherited types include cystine calculi in certain dachshunds and uric acid calculi in male dalmatians
Hypothyroidism	functional disturbance	thyroid gland may function marginally or be absent; symptoms include awkward, slow movement, coarse, dry coat; treatment includes iodine, thyroid preparations
Dermatitis	varied	common symptoms include skin inflammation and loss of hair; causative agents include nutritional deficiencies, bacterial infections, hypothyroidism, allergies, hormone imbalances, and parasites (*e.g.*, fleas, lice, mites, fly larvae, and ticks)
Strychnine poisoning	chemical compound	accidental ingestion of 0.75 milligram of the poison (found in rat poisons) per kilogram (about 2¼ pounds) of body weight may cause death from convulsions and respiratory distress
Glaucoma	hereditary tendency in some breeds	a group of eye diseases in which the retina and optic nerve are damaged; certain breeds have a hereditary tendency for the disease; other breeds develop glaucoma as a result of other eye disorders
Granulomatous colitis	not yet characterized	usually found in boxer dogs; symptoms include bloody diarrhea; severely and chronically affected dogs become emaciated; an infectious agent observed microscopically in the thickened colon has not yet been isolated or characterized
Pancreatitis	unknown	in acute types the gland may be destroyed because of inflammation from unknown causes; an animal that lives may develop diabetes mellitus or be unable to secrete enzymes from pancreas, or both, thus preventing digestion, which increases the appetite and causes progressive weight loss; treatment difficult

suring the amount of certain chemical constituents of the blood or body fluids, determining the presence of toxins (poisons), or examining the urine and feces. Other tests are designed to identify the causative agents of the disease. The removal and examination of tissue or other material from the body (biopsy) is used to diagnose the nature of abnormalities such as tumours. Specific skin tests are used to confirm the diagnoses of various diseases—*e.g.*, tuberculosis and Johne's disease in cattle and glanders in horses (Table 3).

Confirmation of the presence in the blood of abnormal quantities of certain constituents aids in diagnosing certain diseases. Abnormal levels of protein in the blood are associated with some cancers of the bone, such as multiple myeloma in horses and dogs. Animals with diabetes mellitus have a high level of the carbohydrate glucose and the steroid cholesterol in the blood. The combination of an increase in the blood level of cholesterol and a decrease in the level of iodine bound to protein indicates hypothyroidism (underactive thyroid gland). A low level of calcium in the serum component of blood confirms milk fever in lactating dairy cattle. An increase in the activities of certain enzymes (biological catalysts) in the blood indicates liver damage. An increase in the blood level of the bile constituent bilirubin is used as a diagnostic test for hemolytic crisis, a disease in which red blood cells are rapidly destroyed by organisms such as *Babesia* species in dogs and in cattle and *Anaplasma* species in cattle.

Blood tests

The examination of the formed elements of blood, including the oxygen-carrying red blood cells (erythrocytes), the white blood cells (neutrophils, eosinophils, basophils, lymphocytes, and monocytes), and the platelets, which function in blood coagulation, is helpful in diagnosing certain diseases. Examination of the blood cells of cattle may reveal abnormal lymphocytic cells characteristic of leukemia. Low numbers of leucocytes indicate the presence of viral diseases, such as hog cholera and infectious hepatitis in dogs. Neutrophil levels increase in chronic bacterial diseases, such as canine pneumonia and uterine infections in female animals. Elevated monocyte levels occur in chronic granulomatous diseases; *e.g.*, histoplasmosis and tuberculosis. Canine parasitism and allergic skin disorders are characterized by elevated eosinophil levels. Prolonged clotting time may be associated with a deficiency of platelets.

Anemia has many causes. They include hemorrhages from blood loss after injuries; the destruction of red blood cells by the rickettsia *Haemobartonella felis* in cats; incompatible blood transfusions in dogs; the inadequate production of normal red blood cells, which occurs in iron or cobalt deficiency after exposure to radioactive substances; general malnutrition; and contact with substances that depress the activity of bone marrow.

Poisonings occur commonly in animals. Some species are more sensitive to certain poisons than others. Swine develop mercury poisoning if they eat too much grain that has been treated with mercury compounds to retard spoilage. Dogs may be poisoned by the arsenic found in pesticides or by strychnine, which is found in rat poison. Many plants are poisonous if eaten, such as bracken fern, which poisons cattle and horses, and ragwort, which contains a substance poisonous to the liver of cattle.

Examination of an animal's urine may reveal evidence of kidney diseases or diseases of the entire urinary system or a generalized systemic disease. The presence of protein in the urine of dogs indicates acute kidney disease (nephritis). Although constituents of bile normally are found in the urine of dogs, the quantity increases in dogs with the presence of infectious hepatitis, a disease of the liver. The presence of abnormal amounts of the simple carbohydrate glucose and of ketone bodies (organic compounds involved in metabolism) in an animal's urine is used to diagnose diabetes mellitus, a disease in which the pancreas cannot form adequate quantities of a substance (insulin) important in regulating carbohydrate metabolism. The urine of horses with azoturia (excessive quantities of nitrogen-containing compounds in the urine, Table 4) or muscle breakdown may contain a dark-coloured molecule called myoglobin.

The presence of eggs or parts of worms in the excrement of animals suspected of suffering from intestinal parasites, such as roundworms, tapeworms, or flatworms, aids in diagnosis. Feces that are light in colour, have a rancid odour, contain fat, and are poorly formed may indicate the existence of a chronic disease of the pancreas. Clay-coloured fatty feces suggest obstruction of the bile duct, which conveys bile to the intestine during digestion.

Identification of causative agents

The identification of a disease-causing microorganism within an animal enables the veterinarian to choose the best drug for therapy. Agglutination tests, which utilize serum samples of animals and microorganisms suspected of causing a disease, many times confirm the presence of the following bacterial diseases (Table 3): brucellosis in cattle and swine, salmonellosis in swine, leptospirosis in cattle, and actinobacillosis in swine and cattle. Other tests measure the antibodies (specific proteins formed in re-

Table 6: Some Common Diseases of Domestic Cats

name(s) of disease	causative agent	nature of disease
Feline distemper (panleukopenia, infectious enteritis)	virus	the most important viral disease of cats; wildcats and raccoons also susceptible; number of white blood cells decreases; fluid losses cause dehydration; treatment includes replacing fluid and preventing bacterial infections; vaccination advisable in kittens
Feline rhinotracheitis (pneumonitis, coryza, and influenza)	viruses (*e.g., Miyagawanella*)	any upper-respiratory viral infection with eye and nose involvement; common among cats; seldom causes death; similar to severe head cold in man; treatment includes antibiotic drugs
Feline picornavirus pneumonia	virus	symptoms include watery eyes, nasal inflammation; pneumonia; in severe cases, death is preceded by laboured breathing
Lymphocytic leukemia	virus	cancerous lymphocytic cells enter blood from a malignant tumour (lymphosarcoma); lymph nodes may become enlarged; a progressive anemia, an increase in the level of circulating immature lymphocytes, symptoms may precede death; no curative treatment known thus far
Urinary-bladder infection with stones and obstruction (cystitis)	unknown	urethra is obstructed with crystals of magnesium ammonium phosphate and mucus associated with infection of the urinary bladder (cystitis); death occurs if obstruction not relieved; exact cause of syndrome (set of symptoms) unknown but may be related to diet or to a virus
Ringworm (dermatomycosis)	fungus (*Microsporum canis; Trichophyton* species)	kittens most often affected; disease (a scaly, spreading skin condition) may be transmitted to children; some cats may carry the disease and show no clinical signs
Toxoplasmosis	protozoan (*Toxoplasma gondii*)	toxoplasmosis probably occurs frequently in cats; chronic forms affect abdominal organs; organisms may invade nearly any body tissue
Ear mites, ear inflammation (otitis externa)	mite (*Otodectes cynotis*)	acquired through contact of kitten with an infected mother; constant scratching of ears eventually causes raw sores and scabs; treatment includes killing the mites and controlling secondary bacterial infections
Osteodystrophy fibrosa (osteogenesis imperfecta)	nutritional deficiencies	once considered congenital in kittens but probably is nutritional, resulting from a calcium deficiency; lameness, first noted at ten to 15 weeks, may lead to paralysis of rear legs, slow growth, bone fractures, and severe pain; treatment includes a diet with adequate calcium
"Fur ball" disease	physical agent	hair balls accumulate in stomach as a result of constant grooming; surgical removal sometimes necessary

Table 7: Some Common Diseases of Domestic Poultry

name(s) of disease	causative agent	nature of disease
Coccidiosis	protozoans (*e.g., Emieria tenella*)	affected birds show low egg production, poor growth rates, and high mortality; cecal coccidiosis (the cecum is a pouch in the large intestine), a serious disease caused by tenella; many protozoa produce intestinal symptoms
Blackhead (histomoniasis)	protozoan (*Histomonas meleagridis*)	may kill up to 50 percent of a turkey flock; symptoms include droopy wings and damage to liver and cecum; *Heterakis gallinae*, a worm in the turkey cecum, probably transmits the protozoan
Psittacosis (ornithosis)	virus (*Bedsonia*)	affects parakeets, canaries, parrots, pigeons, and other pet birds; symptoms include lack of appetite, ruffled feathers; can be transmitted to man
Avian lymphomatosis	virus	an infectious disease; manifestations include formation of tumours; three forms occur, depending on location of tumours: visceral (internal organs); neural (nerve); and ocular (eye) lymphomatosis
Pullorum disease	bacterium (*Salmonella pullorum*)	affects most species of fowl (chickens); mortality rate high, also reduces egg productivity of mature females; transmitted from egg-producing organs of hen to chick; disease causes hemorrhages throughout body
Digestive diseases in caged birds	functional disturbance	symptoms include slowly emptying, loosely hanging crop (digestive organ); deficiency of grit (particles that aid in digestion) results in poor nutrition; symptoms include obstruction of crop, vomiting, intestinal inflammation, and various liver disorders
Fractured legs in caged birds	physical agent	in canaries, about 70 percent of fractures involve the metatarsus (hind-foot); in budgerigars, 70 percent involve the tibia (hindleg)

sponse to a foreign substance in the body) formed against a disease-causing agent, such as those that cause brucellosis, foot-and-mouth disease, infectious hepatitis in dogs, and fowl pest.

The modern veterinary diagnostic laboratory performs, in addition to the tests mentioned, tests of cells in the bone marrow; specific-organ-function tests (liver, kidney, pancreas, thyroid, adrenal, and pituitary glands); radioisotope tests, tissue biopsies, and histochemical analyses; and tests concerning blood coagulation and body fluids.

Survey of animal diseases

INFECTIOUS AND NONINFECTIOUS DISEASES

Diseases may be either infectious or noninfectious. The term infection, as observed earlier, implies an interaction between two living organisms, called the host and the parasite. Infection is a type of parasitism, which may be defined as the state of existence of one organism (the parasite) at the expense of another (the host). Agents (*e.g.*, certain viruses, bacteria, fungi, protozoans, worms, and arthropods) capable of producing disease are pathogens. The term pathogenicity refers to the ability of a parasite to enter a host and produce disease; the degree of pathogenicity—that is, the ability of an organism to cause infection—is known as virulence. The capacity of a virulent organism to cause infection is influenced both by the characteristics of the organism and by the ability of the host to repel the invasion and to prevent injury. A pathogen may be virulent for one host but not for another. Pneumococcal bacteria, for example, have a low virulence for mice and are not found in them in nature; if introduced experimentally into a mouse, however, the bacteria overwhelm its body defenses and cause death.

Many pathogens (*e.g.*, the bacterium that causes anthrax, Table 3) are able to live outside the animal's body until conditions occur that are favourable for entering and infecting it. Pathogens enter the body in various ways—by penetrating the skin or an eye, by being eaten with food, or by being breathed into the lungs. After their entry into a host, pathogens actively multiply and produce disease by interfering with the functions of specific organs or tissues of the host. Table 3 lists some infectious and parasitic diseases of animals and the causative agents.

Before a disease becomes established in a host, the barrier known as immunity must be overcome. Defense against infection is provided by a number of chemical and mechanical barriers, such as the skin, mucous membranes and secretions, and components of the blood and other body fluids. Antibodies, which are proteins formed in response to a specific substance (called an antigen) recognized by the body as foreign, are another important factor in preventing infection. Immunity among animals varies with species, general health, heredity, environment, and previous contact with a specific pathogen.

As certain bacterial species multiply, they may produce and liberate poisons, called exotoxins, into the tissues; other bacterial pathogens contain toxins, called endotoxins, which produce disease only when liberated at the time of death of the bacterial cell. Some bacteria, such as certain species of *Clostridium* and *Bacillus*, have inactive forms called spores, which may remain viable (*i.e.*, capable of

Table 8: Some Common Diseases of Fish

name(s) of disease	causative agent	nature of disease
Mouth fungus (cotton-wool disease)	bacterium (*Chondrococcus columnaris*)	fungus-like disease inaccurately named, since causative agent is a bacterium; a contagious disease; produces swollen lips, loss of appetite, and a "cotton-wool-like" growth on mouth; treatment utilizes antibiotic drugs
Tailrot	bacteria (*Haemophilus* and *Aeromonas* species)	infection may spread from fin and tail to body and cause death; disease may be controlled by surgery or use of drugs
Dropsy (ascites)	possibly associated with a bacterium (*Aeromonas punctata*)	characterized by accumulation of liquid in internal organs and tissues, inflammation of intestines, and infection of liver; epidemics can occur; most treatment except the antibiotic chloramphenicol unreliable
Red pest of eels	bacteria (*Vibrio anguillarum*; *Aeromonas* species)	*Vibrio* multiplies readily in salty water (1.5 to 3.5 percent) and can cause extensive blood-coloured areas on skin; *Aeromonas* species produce ulcers in the skin
Fish tuberculosis	bacteria (*Mycobacterium* species)	symptoms include loss of appetite, emaciation (leanness), skin defects, blood spots, ulcers, and cysts (on internal organs)
Eye fungus	fungus	infection follows damage to cornea of eye; a typical symptom is a white cotton-wool growth hanging from eye; untreated eye is destroyed within days; untreated fish die
Fish lice	louse (*Argulus* species)	bloodsucking parasites on the surfaces of many fish species
Skin flatworms (flukes)	platyhelminth	small parasites; cause skin colour to fade, blood spots, increased respiratory rate, and debilitation
Bursting of the swim bladder	physical agent	occurs if fish in deep water rise to surface too rapidly; fatal
Air embolism	physical agent	occurs if oxygen content of water is higher than normal, as when water temperature is higher than normal; bubbles of nitrogen gas in blood cause the disease

developing into active organisms) for many years; spores are highly resistant to environmental conditions such as heat, cold, and chemical compounds called disinfectants, which are able to kill many active bacteria.

The term infestation indicates that animals, including spiny-headed worms (Acanthocephala), roundworms (Nematoda), flatworms (Platyhelminthes), and arthropods such as lice, fleas, mites, and ticks, are present in or on the body of a host. An infestation is not necessarily parasitic. Table 3 includes various infestations.

Non-infectious diseases

Noninfectious diseases are not caused by virulent pathogens and are not communicable from one animal to another (see Table 4). They may be caused by hereditary factors or by the environment in which an animal lives. Many metabolic diseases are caused by an unsuitable alteration, sometimes brought about by man, in an animal's genetic constitution or in its environment. Metabolic diseases usually result from a disturbance in the normal balance of the physiological mechanisms that maintain stability, or homeostasis. Examples of metabolic diseases include overproduction or underproduction of hormones, which control specific body processes; nutritional deficiencies; poisoning from such agents as insecticides, fungicides, herbicides, fluorine, and poisonous plants; and inherited deficiencies in the ability to synthesize active forms of specific enzymes, which are the proteins that control the rates of chemical reactions in the body.

Excessive inbreeding (*i.e.*, the mating of related animals) among all domesticated animal species has resulted in an increase in the number of metabolic diseases and an increase in the susceptibility of certain animals to infectious diseases.

ZOONOSES

As stated previously, zoonoses are human diseases acquired from or transmitted to any other vertebrate animal. Zoonotic diseases are common in currently developing countries throughout the world and constitute, with starvation, the major threat to human health. More than 150 such diseases are known; some examples are listed in Table 10.

Types of zoonoses

Zoonoses may be separated into four principal types, depending on the mechanisms of transmission and epidemiology. One type includes the direct zoonoses, such as rabies and brucellosis, which are maintained in nature by one vertebrate species. The transmission cycle of the cyclozoonoses, of which tapeworm infections are an example, requires at least two different vertebrate species. Both vertebrate and invertebrate animals are required as intermediate hosts in the transmission to humans of metazoonoses; arboviral and trypanosomal diseases are good examples of metazoonoses. The cycles of saprozoonoses (for example, histoplasmosis) may require, in addition to vertebrate hosts, specific environmental locations or reservoirs.

Most animals that serve as reservoirs for zoonoses are domesticated and wild animals with which man commonly associates. People in occupations such as veterinary medicine and public health, therefore, have a greater

Table 9: Some Important Diseases of Some Common Laboratory Animals

name or type of disease	causative organism
Mice	
Bacterial diseases	
Mouse pneumonitis	*Miyagawanella broncho-pneumoniae*
Eperythrozoonosis	*Eperythrozoon coccoides*
Mycoplasma infection	*Mycoplasma* species
Pus-producing lesions	*Pseudomonas aeruginosa*
Salmonellosis	*Salmonella*
Pseudotuberculosis	*Pasteurella pseudotuberculosis*
Rat-bite fever	*Streptobacillus moniliformis*
Fungal disease	
Dermatophytoses (ringworm)	*Trichophyton* species
Viral diseases	
Mouse pox (ectromelia)	Ectromelia virus
Poliomyelitis of mice (Theiler's disease)	Theiler's encephalomyelitis virus
EDIM (epizootic diarrhea of infant mice)	EDIM virus
LIVIM (lethal intestinal virus of infant mice)	LIVIM virus
Protozoal diseases	
Coccidiosis	*Eimeria falciformis*
Toxoplasmosis	*Toxoplasma gondii*
Nosematosis (encephalitozoonosis)	*Nosema cuniculi*
Helminthic diseases	
Mouse tapeworm	*Hymenolepis nana*
Rat tapeworm	*Hymenolepis diminuta*
Cat tapeworm	*Taenia taeniaeformis*
Mouse pinworm	*Aspicularis tetraptera*
Pinworm	*Syphacia obvelata*
Liver threadworm	*Capillaria hepatica*
Arthropod diseases	
Sucking louse	*Polyplax serratus*
Mite	*Psorergates simplex*
Rats	
Bacterial and fungal diseases	
Mycoplasma infection (associated with chronic pneumonia)	*Mycoplasma* species
Salmonellosis	*Salmonella* species
Infectious anemia	*Haemobartonella muris*
Pus-forming lesions	*Pseudomonas aeruginosa*
	Pasteurella pneumotropica
Rat-bite fever	*Streptobacillus moniliformis*
	Diplococcus pneumoniae
Leptospirosis (Weil's disease)	*Leptospira* species
	Corynebacterium kutscheri
Dermatophytoses (ringworm; fungus)	*Trichophyton*
Viral diseases	
Pneumonitis	rat virus (Kilham)
Salivary-gland disease	rabula virus
Hemorrhagic encephalitis	hemorrhagic encephalopathy virus (HER)
Rats (cont.)	
Protozoal diseases	
Coccidiosis	*Eimeria nieschulzi*
	Hepatozoon muris
Toxoplasmosis	*Toxoplasma gondii*
Nosematosis (encephalitozoonosis)	*Nosema cuniculi*
Helminthic diseases	
Mouse tapeworm	*Hymenolepis nana*
Rat tapeworm	*Hymenolepis diminuta*
Cat tapeworm	*Taenia taeniaeformis*
Bladder threadworm	*Trichosomoides crassicauda*
Cecal worm	*Heterakis spumosa*
Liver threadworm	*Capillaria hepatica*
Arthropod diseases	
Sucking louse	*Polyplax spinulosa*
Ear mite	*Notoedres notoedres*
Hair mite	*Radifordia ensifera*
Guinea pigs	
Bacterial and fungal diseases	
Streptococcal infections (cervical abscesses and lymphadenitis)	*Streptococcus* species
Salmonellosis	*Salmonella* species
Bronchopneumonia	*Bordatella bronchisepitca*
Respiratory infections	*Klebsiella* species
Dermatophytoses (ringworm, fungus)	*Trichophyton* species
Viral diseases	
Cytomegalic inclusion disease	salivary-gland virus
Lymphocytic choriomeningitis	LCM virus
Protozoal diseases	
Coccidiosis	*Eimeria caviae*
Toxoplasmosis	*Toxoplasma gondii*
Hamsters	
Cytomegalic inclusion disease	salivary-gland virus
Encephalomyocarditis	picorna virus
Wet tail (regional ileitis)	virus, bacterium
Salmonellosis	*Salmonella* (bacterium)
Mouse tapeworm	*Hymenolepsis nana* (helminth)
Rabbits	
Pasteurellosis (snuffles)	*Pasteurella* (bacterium)
Spirochetosis (vent disease)	*Spirochaeta* (bacterium)
Mucoid enteritis	bacterium
Myxomatosis	myxoma virus
Rabbit pox	virus
Coccidiosis	*Eimeria* species (protozoan)
Ear mites	*Notoedres notoedres*
Nosematosis	*Nosema cuniculi*

Table 10: A Partial List of Zoonoses

disease	causative organism	animals principally involved
Viral diseases		
Arbovirus infections		rodents, birds, equines, goats, sheep, monkeys, swine, marsupials
Febrile illnesses	various arboviruses	
Hemorrhagic fever		
Epidemic nephroso-nephritis		
Encephalitis (mosquito-borne and tick-borne)		
Encephalomyocarditis	encephalomyocarditis virus	rodents
Herpes B virus disease	herpes B virus	monkeys
Herpes T (= M) virus infection	herpes T (= M) virus	monkeys
Influenza	influenza virus type A	swine
Lymphocytic choriomeningitis	lymphocytic choriomeningitis virus	mice, dogs, monkeys
Newcastle disease	Newcastle disease virus	chickens
Poxvirus infections		
Buffalopox	buffalopox virus	buffalo
Camelpox	camelpox virus	camels
Cowpox	cowpox or vaccinia virus	cattle
Orf (contagious ecthyma)	contagious ecthyma virus	sheep and goats
Paravaccinia (milkers' nodules)	paravaccinia virus	cattle
Yaba disease	yaba virus	monkeys
Rabies	rabies virus	carnivores, bats, and other wild animals
Sendai virus disease	sendai virus	swine, rodents
Cat-scratch disease	cat-scratch virus	cats
Rickettsial diseases		
Flea-borne		
Murine (endemic typhus)	*Rickettsia mooseri*	rats, mice
Mite-borne		
Rickettsial pox	*R. akari*	mice
Scrub typhus (tsutsugamushi)	*R. tsutsugamushi*	rodents
Tick-borne		
(North) Queensland tick typhus	*R. australis*	bandicoots, rodents
Spotted fever (including Rocky Mountain, Brazilian, and Colombian spotted fevers)	*R. rickettsii*	dogs, rodents, and other animals
Fièvre boutonneuse	*R. conorii*	dogs, rodents
Kenya typhus		
South African tick typhus		
Indian tick typhus		
North Asian tick-borne rickettsiosis	*R. sibericus*	rodents

disease	causative organism	animals principally involved
Rickettsial diseases (cont.)		
Q fever	*Coxiella burnetii*	cattle, sheep, goats, and other domesticated and wild mammals, birds
Bedsonia infection		
Psittacosis (ornithosis)	Psittacosis (P.L.T.) group (*Bedsonia*)	psittacines and other birds
Bacterial diseases		
Anthrax	*Bacillus anthracis*	ruminants, equines, swine
Brucellosis	*Brucella abortus, B. suis, B. melitensis*	cattle, swine, goats, sheep, horses
Enterobacterial infections		
Arizona infections	*Arizona* species	poultry, swine, dogs
Colibacillosis	*E. coli*	poultry, swine, dogs
Salmonellosis	*Salmonella* species	mammals and birds
Erysipeloid	*Erysipelothrix rhusiopathiae*	swine, poultry, fish
Glanders	*Actinobacillus mallei*	equines
Leptospirosis	*Leptospira interrogans*	rodents, dogs, swine, cattle, bandicoots
Listeriosis	*Listeria monocytogenes*	rodents, sheep, cattle, swine
Melioidosis	*Pseudomonas pseudomallei*	rodents, sheep, cattle, swine
Pasteurellosis	*Pasteurella multocida, P. haemolytica*	mammals, birds
Plague	*Pasteurella pestis*	rodents
Pseudotuberculosis	*Pasteurella pseudotuberculosis*	rodents, cats, fowls
Rat-bite fever	*Spirillum minus, Streptobacillus moniliformis*	rodents
Relapsing fever (tick-borne)	*Borrelia* species	rodents
Staphylococcosis	*Staphylococcus aureus*	cattle, dogs, occasionally other animals
Streptococcosis	*Streptococcus* species	mammals
Tuberculosis	*Mycobacterium tuberculosis var. hominis*	dogs, swine, monkeys
	M. bovis	cattle, goats, swine, cats
	M. avium	poultry, swine, cattle
Tularaemia	*Pasteurella tularensis*	rabbits, hares, sheep, wild rodents
Vibriosis	*Vibrio fetus; V. parahaemolyticus*	cattle, sheep fish
Fungal disease		
Dermatophytosis		
Ringworm, favus	*Microsporum* species *Trichophyton* species	cats, dogs, horses horses, cattle, poultry, small mammals

exposure to zoonoses than do those in occupations less closely concerned with animals.

In addition to the numerous human diseases spread by contact with the parasitic worm helminth and by contact with arthropods (see Table 10), many diseases are transmitted by the bites and venom of certain animals; poisonous or diseased food animals also transmit diseases. Dog bites may seriously injure tissues and also can transmit bacterial infections and rabies, a disease of viral origin. The bite of a diseased rat may transmit any of several diseases to man, including plague, salmonellosis, leptospirosis, and rat-bite fevers. Cat-scratch disease may be transmitted through cat bites, and the deadly herpes B virus can spread by monkey bites. The bites of venomous snakes and fish account for considerable human discomfort and death. About 200 of the 2,500 known species of snakes can cause human disease. One estimate for snakebite deaths worldwide is 30,000 to 40,000 per year, the vast majority of them in Asia. Poisonous wild animals inadvertently used for food include animals harbouring the anthrax bacillus and those containing the causative agents of salmonellosis, trichinosis, and fish-tapeworm infection. The flesh of various types of fish is toxic to man. Japanese puffers, for example, contain the poisonous chemical compound tetrodotoxin; scombroid fish harbour *Proteus morganii,* which causes gastrointestinal diseases; and mullet and surmullet can cause nervous disturbances.

Control of zoonoses

Approaches to the control of zoonoses differ according to the type under consideration. Because the majority of direct zoonoses and cyclozoonoses and some saprozoonoses are most effectively controlled by techniques involving the animal host, methods used to combat these diseases are almost entirely the responsibility of veterinary medicine. A good example is the elimination of stray dogs, for they are an important factor in the control of zoonoses such as rabies, hydatid disease, and visceral larva migrans. In addition, the control of diseases such as brucellosis and tuberculosis in cattle involves a combination of methods—mass immunization, diagnosis, slaughter of infected animals, environmental disinfection, and quarantine. Several supportive measures for the control of disease are useful in some cases. Air-sanitation measures are helpful in direct zoonoses in which human illness is spread by droplets or dust, and zoonotic infections that are spread through a fluid medium, such as water or milk, sometimes can be controlled. Heat, cold, and irradiation are effective in killing the immature forms of *Trichinella spiralis,* the causative agent of trichinosis, in meat; and certain antibiotic drugs help to prevent deterioration of food.

The control of metazoonoses may be directed at the infected vertebrate hosts, at the infected invertebrate host, or at both. Particularly effective in this instance has been the use of chemical insecticides to attack the invertebrate carriers of specific infections, even though several difficulties have been encountered—for example, the inaccessibility of the invertebrate to the chemicals, which occurs with organisms that breed in swiftly flowing waters or in dense vegetation, and the development of insecticide re-

Table 10: A Partial List of Zoonoses (continued)

disease	causative organism	animals principally involved
Protozoal diseases		
Amebiasis	*Entamoeba histolytica*	dogs, lower primates
Balantidiasis	*Balantidium coli*	swine
Coccidiosis	*Isospora* species	dogs
Leishmaniasis		
Kala Azar	*Leishmania donovani*	dogs
Oriental sore	*Leishmania tropica*	dogs, rodents
American	*Leishmania* species	dogs, wild mammals
Malaria	*Plasmodium knowlesi*	monkeys
	Plasmodium simium	monkeys
	Plasmodium cynomolgi	monkeys
Pneumocystis infection	*Pneumocystis carinii*	dogs
Toxoplasmosis	*Toxoplasma gondii*	mammals, birds
Trypanosomiasis	*Trypanosoma cruzi*	dogs, small mammals
	Trypanosoma rangeli	
	Trypanosoma rhodesiense	antelope, cattle
Platyhelminthic diseases		
Trematode (fluke) diseases		
Amphistomiasis	*Gastrodiscoides hominis*	swine
Cercarial dermatitis	*Schistosoma* species	birds, mammals
Clonorchiasis	*Clonorchis sinensis*	dogs, cats, swine, wild mammals, fish
Dicrocoeliasis	*Dicrocoelium* species	ruminants
Echinostomiasis	*Echinostoma ilocanum*	cats, dogs, rodents
	Echinostoma species	
Fascioliasis	*Fasciola hepatica, F. gigantica*	ruminants
Fasciolopsiasis	*Fasciolopsis buski*	swine, dogs
Heterophyiasis	*Heterophyes heterophyes* (and other heterophids)	cats, dogs, fish
Metagonimiasis	*Metagonimus yokogawai*	cats, dogs, fish
Opisthorchiasis	*Opisthorchis felineus*	cats, dogs
	Opisthorchis viverini, other species	wildlife, fish
Paragonimiasis	*Paragonimus westermani*, other species	cats, dogs, wildlife
Schistosomiasis	*Schistosoma japonicum*	wild and domestic mammals
	S. mansoni	baboons, rodents
	S. mattheei, occasionally other species	cattle, sheep, antelopes
Cestode (tapeworm) diseases		
Bertiella infection	*Bertiella studeri*	primates
Diphyllobothriasis	*Diphyllobothrium latum*	fish, carnivores
Dipylidiasis	*Dipylidium caninum*	dogs, cats
Echinococcosis	*Echinococcus granulosus*	dogs, wild carnivores, domestic and wild ungulates
	E. multilocularis	foxes, dogs, rodents
Hymenolepiasis	*Hymenolepis diminuta, H. nana*	rats, mice
Inermicapsifer infection	*Inermicapsifer madagascarensis*	rodents

disease	causative organism	animals principally involved
Platyhelminthic diseases (cont.)		
Sparganosis	Pseudophyllidea tapeworms	mice, carnivores including cats, and other vertebrates
Taeniasis, cysticercosis, and coenuriasis	*Taenia saginata*	cattle
	Taenia solium	swine
	Multiceps multiceps	sheep, dogs
Nematode diseases		
Ancylostomiasis	*Ancylostoma ceylanicum*, other species	dogs
Ascariasis	*Ascaris suum*	swine
Capillariasis	*Capillaria hepatica*	rodents
Dracunculosis	*Dracunculus medinensis*	dogs, other carnivores
Filariasis	*Brugia malayi*	primates, other mammals
	Dirofilaria species, occasionally other species	cats, dogs, other mammals
Larva migrans	*Ancylostoma braziliense*, other species	cats, dogs
	Angiostrongylus cantonensis	rats
	Anisakis species	fish
	Gnathostoma spinigerum	cats, dogs, other vertebrates
	Toxocara canis, other ascarid species	dogs, other vertebrates
Oesophagostomiasis	*Oesophagostomum apiostomum*	primates
Strongyloidiasis	*Strongyloides stercoralis*, occasionally other species	dogs, primates
Ternidens infection	*Ternidens deminutus*	primates
Trichinosis	*Trichinella spiralis*	swine, rodents, wild carnivores, marine mammals
Trichostrongylosis	*Trichostrongylus colubriformis*, occasionally other species	ruminants
Arthropod diseases		
Acariasis	*Sarcoptes* species	domesticated animals
Tunga infections	*Tunga penetrans*	domesticated and wild mammals
Myiasis	*Cochliomyia, Cordylobia, Dermatobia, Gastrophilus, Hypoderma, Oestrus*, and other genera	mammals
Pentastomid infections (including Halzoun)	*Linguatula* species, *Armillifer* species, *Porocephalus* species	dogs, snakes, and other vertebrates

sistance by the organisms. Insecticides are used to destroy the mosquitoes that spread malaria (*Anopheles*). Mechanical filters placed across irrigation ditches help to prevent the dissemination of the snails that transmit *Schistosoma mansoni,* a parasitic flatworm.

DISEASE PREVENTION, CONTROL, AND ERADICATION

Prevention is the first line of defense against disease. At least four preventive techniques are available for use in the prevention of disease in an animal population. One is the exclusion of causative agents of disease from specific geographic areas, or quarantine. A second preventive tool utilizes control methods such as immunization, environmental control, and chemical agents to protect specific animal populations from endemic diseases, diseases normally present in an area. The third preventive measure concerns the mass education of people about disease prevention. Finally, early diagnosis of illness among members of an animal population is important so that disease manifestations do not become too severe and so that affected animals can be more easily managed and treated.

Quarantine—the restriction of movement of animals suffering from or exposed to infections such as bluetongue and scrapie (in sheep), foot-and-mouth disease (in cattle), and rabies (in dogs)—is one of the oldest tools of preventive medicine. It was applied to domesticated animals as early as Roman times. The establishment of international livestock quarantine in the United States in 1890 provided for the holding of all imported cattle, sheep, and swine at the port of entry for 90, 15, and 15 days, respectively. In this way, such diseases as Nairobi sheep disease, surra, and infections caused by *Brucella melitensis* were eliminated or excluded from the United States, but international quarantine barriers did not prevent the entry of bluetongue, scrapie, and the tick *Rhipicephalus evertsi,* which is a carrier for several animal diseases. On the other hand, long-term quarantine of all dogs entering Great Britain has been effective since its initiation in 1919 (the quarantine also includes cats). It is possible that aircraft may pose new problems regarding livestock-disease quarantine since many disease carriers (*e.g.,* insects and viruses) may be accidentally brought by plane into a country.

Contact through travel

Mass immunization as a preventive technique has the advantage of allowing the resistant animal freedom of movement, unlike environmental control, in which the animal is confined to the controlled area; immunization may, however, provide only short-lived and partial protection. Mass-inoculation techniques against diseases such as Newcastle disease in chickens and distemper in mink and dogs have been successful. Animal diseases have been prevented by methods involving environmental control, including the maintenance of safe water supplies, the hygienic disposal of animal excrement, air sanitation, pest control, and the improvement of animal housing. One specific environmental program, called the portable-calf-pen system, involves routine movement of the pens to avoid a concentration of specific pathogens in them. Other programs involve the utilization of automatic and sani-

Table 11: Animal Diseases Usually Confined to Certain Regions of the World

name(s) of disease	animal(s) affected	causative organism	distribution	nature of disease
African horse sickness (AHS), equine plague, pestis equorum, perdesiekte	primarily horses, donkeys, mules (occasionally zebras and dogs)	AHS virus	primarily Africa and Middle East; occasionally India, Pakistan	a seasonal disease occurring in late summer; acute form, sometimes fatal within five days, involves excessive fluid in lungs; symptoms of other forms include accumulation of fluid in body cavities
African swine fever (ASF), warthog disease, Montgomery's disease	swine	ASF virus	primarily Kenya and South Africa; occasionally Europe	highly contagious; usually fatal; resembles hog cholera in clinical manifestations (high fever, weakness in hindlegs, and hemorrhages throughout body) but can be distinguished by laboratory tests and isolation of the virus
Contagious pleuropneumonia, lung plague	cattle, buffalo, yaks, sheep, goats	*Mycoplasma mycoides*	Africa, Australia, Asia, Europe	transmitted by direct animal contact or by contaminated objects; an acute disease producing pneumonia and inflammation of the lung lining; vaccines ineffective because different strains of the organism occur throughout the world
East coast fever, theileriasis, Rhodesian red water or tick fever	cattle, African and Indian water buffalo	protozoan (*Theileria parva*)	Central Africa; East Africa	usually fatal; transmitted by three ticks containing pathogen; symptoms include high fever, swelling of lymph glands; not yet prevented effectively by vaccination
Foot-and-mouth disease (FMD), aphthous fever, aftosa	cattle, swine, sheep, goats	FMD virus	worldwide except North America, Central America, New Zealand	symptoms include high fevers, drool from mouth, where vesicles and ulcers form, and lameness; causes great economic losses throughout world; effective vaccines available
Fowl plague, fowl pest	birds, including chickens and turkeys	fowl-plague virus	Europe, Middle and Far East, Argentina, Japan	may cause no apparent symptoms; apparent symptoms include lack of appetite, swollen head, laboured breathing, and hemorrhaging
Heartwater, drunk bull sickness	cattle, sheep, goats	rickettsia (*Cowdria ruminantium*)	Africa (southern half); Madagascar	disease has acute and mild forms; symptoms include water in the membrane around heart and in the lung cavity, hemorrhages, and twitching
Louping ill (LI), infectious encephalomyelitis in sheep, trembling ill	primarily sheep (also cattle and man)	LI virus	British Isles, Czechoslovakia, U.S.S.R.	transmitted by bite of sheep tick; characterized by fever, dullness followed by excitement, muscular spasms, leaping gait, convulsions, and death
Nagana, tsetse disease, trypanosomiasis	most domesticated animals	protozoan (*Trypanosoma* species)	Africa	may be acute or inapparent; symptoms may include anemia resulting from red-blood-cell destruction; pathogen transmitted by tsetse fly (over 20 species of *Glossina*); prevents effective cattle production in nearly all of West Africa
Rift Valley fever (RVF), infectious enzootic hepatitis	cattle, sheep (occasionally man)	RVF virus	Central and South Africa	spread by bloodsucking insects associated with wild animals; symptoms include abdominal pain resulting from liver damage; young animals usually die; mature ones may recover
Rinderpest, cattle plague	cattle, sheep, goats, wild ruminants; yaks, caribou, gazelles, deer	rinderpest virus	primarily Asia, Africa, Philippines; rarely Europe	rapidly fatal; symptoms include fluid losses (dehydration) from diarrhea caused by massive pathological changes (*e.g.*, hemorrhages, ulcers) in intestinal tract
Surra	primarily in camels and horses; many animals susceptible	protozoan (*Trypanosoma evansi*)	primarily Far East (*e.g.*, China), India, Near East (*e.g.*, Iran); North Africa	transmitted by bloodsucking flies and mosquitoes; symptoms include anemia, loss of weight, large swellings in limbs, abdomen, and sex organs
Teschen disease, swine encephalomyelitis, porcine poliomyelitis	swine	Teschen virus	primarily Europe	symptoms include prostration, immobilization, nervous tremors, convulsions, paralysis of legs

tary watering and feeding equipment and buildings with environmental controls. The use of chemical compounds to prevent illness (chemoprophylaxis) includes a variety of pesticides, which are used to kill insects that transmit diseases, and substances either used internally or applied to the animal's body to prevent the transmission or the development of a disease. An example is the use of sulfonamide drugs in the drinking water of poultry to prevent coccidiosis (see Table 7). Environmental-control methods in the poultry industry have resulted in the most efficient means of poultry production developed thus far.

The early detection of a disease in a population of animals—a herd of cattle, for example—is particularly useful in controlling certain chronic infectious diseases, such as mastitis, brucellosis, and tuberculosis, as well as certain noninfectious diseases such as bloat. Laboratory tests—the agglutination test in pullorum disease, the tuberculin skin test for tuberculosis, the examination of feces for eggs of specific parasites, the physical and chemical tests performed on milk to diagnose bovine mastitis—are used for the early detection of diseases in an animal population.

Methods of disease control and eradication have been successful in various countries. In the United States, for example, the test-and-slaughter technique, in which simple tests are used to confirm the existence of diseased animals that are then slaughtered, has been of great value in controlling infectious and hereditary diseases, including dourine, a venereal disease in horses, fowl plague, and foot-and-mouth disease in cattle and deer. Bovine tuberculosis has been eliminated from Denmark, Finland, and The Netherlands and reduced to a low level in various other countries, including Great Britain, Japan, the United States, and Canada, by the test-and-slaughter method. Many infectious diseases have been eradicated from Great Britain—sheep pox, rinderpest, pleuropneumonia, glanders, and rabies. Diseases eliminated from Australia by a combination of methods—control of agents that carry disease, the test-and-slaughter technique, the use of chemical agents, and, more recently, biological control—include hog cholera, rinderpest, scrapie, glanders, surra, rabies, and foot-and-mouth disease.

In biological control, enemies of the agents that transmit the disease, enemies of the reservoir host, or a specific parasite are introduced into the environment. If a natural enemy of the tsetse fly could be found, for example, African sleeping sickness in man and trypanosomiasis in cattle could be controlled in West Africa. Successful biological control of the European-rabbit population in Australia has been accomplished through the use of the myxomatosis virus, which is transmitted by mosquitoes and causes the

formation of malignant tumours. Although the Brazilian white rabbit is relatively unaffected by the virus, it causes rapid death in the European rabbit. The elimination of the European rabbit in France by the virus was accompanied by a decrease in tick-borne typhus in people, suggesting that the rabbit may be a significant intermediate host for the causative agent, *Rickettsia conorii.* Screwworms, an immature form of the fly *Cochliomyia hominivorax,* have been eradicated in the United States by the release of more than 3,000,000,000 sterilized males.

Disease control and elimination programs require many sophisticated techniques, in addition to diagnosis and the slaughter of affected animals. They include: the control of insects known to transmit diseases; the cooperation of animal owners; the development through research of new diagnostic tests for use on large populations; the eradication of animal species from areas in which they are known to transmit disease; sterilization of strains of animals known to carry inheritable metabolic diseases; and effective meat inspection.

(C.E.Co./Ed.)

DISEASES OF PLANTS

All species of plants, wild and cultivated alike, are subject to disease. Although each species is susceptible to characteristic diseases, these are, in each case, relatively few in number. The occurrence and prevalence of plant diseases vary from season to season, depending on the presence of the pathogen, environmental conditions, and the crops and varieties grown. Some plant varieties are particularly subject to outbreaks of diseases; others are more resistant to them.

General considerations

NATURE AND IMPORTANCE OF PLANT DISEASES

Economic impact of plant diseases

Plant diseases are known from times preceding the earliest writings. Fossil evidence indicates that plants were affected by disease 250 million years ago. The Bible and other early writings mention diseases, such as rusts, mildews, blights, and blast, that have caused famine and other drastic changes in the economy of nations since the dawn of recorded history. Other plant disease outbreaks with similar far-reaching effects in more recent times include late blight of potato in Ireland (1845–60); powdery and downy mildews of grape in France (1851 and 1878); coffee rust in Ceylon (starting in the 1870s); *Fusarium* wilts of cotton and flax; southern bacterial wilt of tobacco (early 1900s); Sigatoka leaf spot and Panama disease of banana in Central America (1900–65); black stem rust of wheat (1916, 1935, 1953–54); and southern corn leaf blight (1970) in the United States.

Loss of crops from plant diseases may result in hunger and starvation, especially in less developed countries where access to disease-control methods is limited and annual losses of 30 to 50 percent are common for major crops. In some years, losses are much greater, producing catastrophic results for those who depend on the crop for food. Major disease outbreaks among food crops have led to famines and mass migrations throughout history. The devastating outbreak of late blight of potato (*Phytophthora infestans*) that began in Europe in 1845 and brought about the Irish famine caused starvation, death, and mass migration of the Irish population. Of a population of eight million, approximately one million (about 12.5 percent) died of starvation and 1.5 million (almost 19 percent) emigrated, mostly to the United States, as refugees from the destructive blight. This fungus thus had a tremendous influence on the economic, political, and cultural development in Europe and the United States. During World War I, late blight damage to the potato crop in Germany may have helped end the war.

Losses from plant diseases also can have a significant economic impact, causing a reduction in income for crop producers and distributors and higher prices for consumers. In 1993 the United States lost more than one million acres (405,000 hectares) of crops to disease. More than 800,000 acres of wheat succumbed to disease, exacting a monetary loss in the millions of dollars.

Diseases—a normal part of nature. Plant diseases are a normal part of nature and one of many ecological factors that help keep the hundreds of thousands of living plants and animals in balance with one another. Humans have carefully selected and cultivated plants for food, clothing, shelter, fibre, and beauty for thousands of years. Disease is just one of many hazards that must be considered when plants are taken out of their natural environment and grown in pure stands under what are often abnormal conditions.

Many valuable crop and ornamental plants are very susceptible to disease and would have difficulty surviving in nature without human intervention. Cultivated plants are often more susceptible to disease than are their wild relatives. This is because large numbers of the same species or variety, having a uniform genetic background, are grown close together, sometimes over many thousands of square kilometres. A pathogen may spread rapidly under these conditions.

Definitions of plant disease. In general, a plant becomes diseased when it is continuously disturbed by some causal agent that results in an abnormal physiological process that disrupts the plant's normal structure, growth, function, or other activities. This interference with one or more of a plant's essential physiological or biochemical systems elicits characteristic pathological conditions or symptoms.

Plant diseases can be broadly classified according to the nature of their primary causal agent, either infectious or noninfectious. Infectious plant diseases are caused by a pathogenic organism such as a fungus, bacterium, mycoplasma, virus, viroid, nematode, or parasitic flowering plant. An infectious agent is capable of reproducing within or on its host and spreading from one susceptible host to another. Noninfectious plant diseases are caused by unfavourable growing conditions, including extremes of temperature, disadvantageous relationships between moisture and oxygen, toxic substances in the soil or atmosphere, and an excess or deficiency of an essential mineral. Because noninfectious causal agents are not organisms capable of reproducing within a host, they are not transmissible.

In nature, plants may be affected by more than one disease-causing agent at a time. A plant that must contend with a nutrient deficiency or an imbalance between soil moisture and oxygen is often more susceptible to infection by a pathogen; a plant infected by one pathogen is often prone to invasion by secondary pathogens. The combination of all disease-causing agents that affect a plant make up the disease complex. Knowledge of normal growth habits, varietal characteristics, and normal variability of plants within a species—as these relate to the conditions under which the plants are growing—is required for a disease to be recognized.

The study of plant diseases is called plant pathology. Pathology is derived from the two Greek words *pathos* (suffering, disease) and *logos* (discourse, study). Plant pathology thus means a study of plant diseases.

DISEASE DEVELOPMENT AND TRANSMISSION

Pathogenesis and saprogenesis. Pathogenesis is the stage of disease in which the pathogen is in intimate association with living host tissue. Three fairly distinct stages are involved:

1. Inoculation: transfer of the pathogen to the infection court, or area in which invasion of the plant occurs (the infection court may be the unbroken plant surface, a variety of wounds, or natural openings—*e.g.,* stomates [microscopic pores in leaf surfaces], hydathodes [stomatelike openings that secrete water], or lenticels [small openings in tree bark])
2. Incubation: the period of time between the arrival of the pathogen in the infection court and the appearance of symptoms

3. Infection: the appearance of disease symptoms accompanied by the establishment and spread of the pathogen.

One of the important characteristics of pathogenic organisms, in terms of their ability to infect, is virulence. Many different properties of a pathogen contribute to its ability to spread through and to destroy the tissue. Among these virulence factors are toxins that kill cells, enzymes that destroy cell walls, extracellular polysaccharides that block the passage of fluid through the plant system, and substances that interfere with normal cell growth. Not all pathogenic species are equal in virulence—that is, they do not produce the same amounts of the substances that contribute to the invasion and destruction of plant tissue. Also, not all virulence factors are operative in a particular disease. For example, toxins that kill cells are important in necrotic diseases, and enzymes that destroy cell walls play a significant role in soft rot diseases.

Many pathogens, especially among the bacteria and fungi, spend part of their life cycles as pathogens and the remainder as saprophytes.

Saprogenesis is the part of the pathogen's life cycle when it is not in vital association with living host tissue and either continues to grow in dead host tissue or becomes dormant. During this stage, some fungi produce their sexual fruiting bodies; the apple scab (*Venturia inaequalis*), for example, produces perithecia, flask-shaped spore-producing structures, in fallen apple leaves. Other fungi produce compact resting bodies, such as the sclerotia formed by certain root- and stem-rotting fungi (*Rhizoctonia solani* and *Sclerotinia sclerotiorum*) or the ergot fungus (*Claviceps purpurea*). These resting bodies, which are resistant to extremes in temperature and moisture, enable the pathogen to survive for months or years in soil and plant debris in the absence of a living host.

Epiphytotics. When the number of individuals a disease affects increases dramatically, it is said to have become epidemic (meaning "on or among people"). A more precise term when speaking of plants, however, is epiphytotic ("on plants"); for animals, the corresponding term is epizootic. In contrast, endemic (enphytotic) diseases occur at relatively constant levels in the same area each year and generally cause little concern.

Epiphytotics affect a high percentage of the host plant population, sometimes across a wide area. They may be mild or destructive and local or regional in occurrence.

Factors leading to epiphytotics

Epiphytotics result from various combinations of factors, including the right combination of climatic conditions. An epiphytotic may occur when a pathogen is introduced into an area in which it had not previously existed. Examples of this condition include the downy mildews (*Sclerospora* species) and rusts (*Puccinia* species) of corn in Africa during the 1950s, the introduction of the coffee rust fungus into Brazil in the 1960s, and the entrance of the chestnut blight (*Endothia parasitica*) into the United States shortly after 1900. Also, when new plant varieties are produced by plant breeders without regard for all enphytotic diseases that occur in the same area to some extent each year (but which are normally of minor importance), some of these varieties may prove very susceptible to previously unimportant pathogens. Examples of this situation include the development of oat varieties with Victoria parentage, which, although highly resistant to rusts (*Puccinia graminis avenae* and *P. coronata avenae*) and smuts (*Ustilago avenae, U. kolleri*), proved very susceptible to *Helminthosporium* blight (*H. victoriae*), formerly a minor disease of grasses. The destructiveness of this disease resulted in a major shift of oat varieties on 50 million acres in the United States in the mid-1940s. Corn (maize) with male-sterile cytoplasm (*i.e.,* plants with tassels that do not extrude anthers or pollen), grown on 60 million acres in the United States, was attacked in 1970 by a virulent new race of the southern corn leaf blight fungus (*Helminthosporium maydis* race T), resulting in a loss of about 700 million bushels of corn. More recently the new *Helminthosporium* race was widely disseminated and was reported from most continents. Finally, epiphytotics may occur when host plants are cultivated in large acreages where previously little or no land was devoted to that crop.

Epiphytotics may occur in cycles. When a plant disease first appears in a new area, it may grow rapidly to epiphytotic proportions. In time, the disease wanes, and, unless the host species has been completely wiped out, the disease subsides to a low level of incidence and becomes enphytotic. This balance may change dramatically by conditions that favour a renewed epiphytotic. Among such conditions are weather (primarily temperature and moisture), which may be very favourable for multiplication, spread, and infection by the pathogen; introduction of a new and more susceptible host; development of a very aggressive race of the pathogen; and changes in cultural practices that create a more favourable environment for the pathogen.

Environmental factors affecting disease development. Important environmental factors that may affect development of plant diseases and determine whether they become epiphytotic include temperature, relative humidity, soil moisture, soil pH, soil type, and soil fertility.

Temperature. Each pathogen has an optimum temperature for growth. In addition, different growth stages of the fungus, such as the production of spores (reproductive units), their germination, and the growth of the mycelium (the filamentous main fungus body), may have slightly different optimum temperatures. Storage temperatures for certain fruits, vegetables, and nursery stock are manipulated to control fungi and bacteria that cause storage decay, provided the temperature does not change the quality of the products. Little, except limited frost protection, can be done to control air temperature in fields, but greenhouse temperatures can be regulated to check disease development.

Disease forecasting

Knowledge of optimum temperatures, usually combined with optimum moisture conditions, permits forecasting, with a high degree of accuracy, the development of such diseases as blue mold of tobacco (*Peronospora tabacina*), downy mildews of vine crops (*Pseudoperonospora cubensis*) and lima beans (*Phytophthora phaseoli*), late blight of potato and tomato (*Phytophthora infestans*), leaf spot of sugar beets (*Cercospora beticola*), and leaf rust of wheat (*Puccinia recondita tritici*). Effects of temperature may mask symptoms of certain viral and mycoplasmal diseases, however, making them more difficult to detect.

Relative humidity. Relative humidity is very critical in fungal spore germination and the development of storage rots. *Rhizopus* soft rot of sweet potato (*Rhizopus stolonifer*) is an example of a storage disease that does not develop if relative humidity is maintained at 85 to 90 percent, even if the storage temperature is optimum for growth of the pathogen. Under these conditions, the sweet potato root produces suberized (corky) tissues that wall off the *Rhizopus* fungus.

High humidity favours development of the great majority of leaf and fruit diseases caused by fungi and bacteria. Moisture is generally needed for fungal spore germination, the multiplication and penetration of bacteria, and the initiation of infection. Germination of powdery mildew spores occurs best at 90 to 95 percent relative humidity. Diseases in greenhouse crops—such as leaf mold of tomato (*Cladosporium fulvum*) and decay of flowers, leaves, stems, and seedlings of flowering plants, caused by *Botrytis* species—are controlled by lowering air humidity or by avoiding spraying plants with water.

Soil moisture. High or low soil moisture may be a limiting factor in the development of certain root rot diseases. High soil-moisture levels favour development of destructive water mold fungi, such as species of *Aphanomyces, Pythium,* and *Phytophthora.* Excessive watering of houseplants is a common problem. Overwatering, by decreasing oxygen and raising carbon dioxide levels in the soil, makes roots more susceptible to root-rotting organisms.

Diseases such as take-all of cereals (*Ophiobolus graminis*); charcoal rot of corn, sorghum, and soybean (*Macrophomina phaseoli*); common scab of potato (*Streptomyces scabies*); and onion white rot (*Sclerotium cepivorum*) are most severe under low soil-moisture levels.

Soil pH. Soil pH, a measure of acidity or alkalinity, markedly influences a few diseases, such as common scab of potato and clubroot of crucifers (*Plasmodiophora brassicae*). Growth of the potato scab organism is suppressed

Controlling acidity of soils

at a pH of 5.2 or slightly below (pH 7 is neutral; numbers below 7 indicate acidity, and those above 7 indicate alkalinity). Scab is not normally a problem when the natural soil pH is about 5.2. Some farmers add sulfur to their potato soil to keep the pH about 5.0. Clubroot of crucifers (members of the mustard family, including cabbage, cauliflower, and turnips), on the other hand, can usually be controlled by thoroughly mixing lime into the soil until the pH becomes 7.2 or higher.

Soil type. Certain pathogens are favoured by loam soils and others by clay soils. *Phymatotrichum* root rot attacks cotton and some 2,000 other plants in the southwestern United States. This fungus is serious only in black alkaline soils—pH 7.3 or above—that are low in organic matter. *Fusarium* wilt disease, which attacks a wide range of cultivated plants, causes more damage in lighter and higher (topographically) soils. Nematodes are also most damaging in lighter soils that warm up quickly.

Soil fertility. Greenhouse and field experiments have shown that raising or lowering the levels of certain nutrient elements required by plants frequently influences the development of some infectious diseases—for example, fire blight of apple and pear, stalk rots of corn and sorghum, *Botrytis* blights, *Septoria* diseases, powdery mildew of wheat, and northern leaf blight of corn. These diseases and many others are more destructive after application of excessive amounts of nitrogen fertilizer. This condition can often be counteracted by adding adequate amounts of potash, a fertilizer containing potassium.

Requirements for disease development. Infectious disease cannot develop if any one of the following three basic conditions is lacking: (1) the proper environment, the most important environmental factors being the amount and frequency of rains or heavy dews, the relative humidity, and the air and soil temperatures, (2) the presence of a virulent pathogen, and (3) a susceptible host. Effective disease-control measures are aimed at breaking this environment-pathogen-host triangle. Loss resulting from disease is reduced, for example, if the host can be made more resistant or immune through such techniques as plant breeding or genetic engineering. In addition, the environment can be made less favourable for invasion by the pathogen and more favourable for the growth of the host plant. Finally, the pathogen can be killed or prevented from reaching the host. These basic methods of control can be divided into a number of cultural, chemical, and biological practices to help control the disease.

The triangle of environment, pathogen, and host

DIAGNOSIS OF PLANT DISEASES

Rapid and accurate diagnosis of disease is necessary before proper control measures can be suggested. It is the first step in the study of any disease. Diagnosis is largely based on characteristic symptoms (Table 12) expressed by the diseased plant. Identification of the pathogen (by "signs," see Table 13) is also essential to diagnosis.

Three steps involved in diagnosis include careful observation and classification of the facts, evaluation of the facts, and a logical decision as to the cause.

Variable factors affecting diagnosis. A skilled diagnostician must know the normal appearance of an affected plant species, its local air and soil environment, the cultural conditions under which it is growing, the pathogens described for the area, and the disease-developing potential of the pathogen. Diagnosis is best done in the presence of the growing plant. Disease is suspected when, for example, part or all of a plant begins to die. Disease also is indicated when blossoms, leaves, stems, roots, or other plant parts appear abnormal—*i.e.,* misshapen, curled, discoloured, overdeveloped, or underdeveloped. Diseased plants also often fail to respond normally to fertilizing, watering, pruning, insect and mite control, or other recommended practices.

Conditions that produce symptoms similar to disease

Conditions other than infection with a pathogen, however, may produce similar or identical symptoms. Some of these have been described, but numerous other conditions must be considered as well when plants are adversely affected. For example, an affected plant may not be adapted to the area in which it is growing. It may not be able to withstand the extremes in soil moisture, temperature, wind, light, or humidity of the local situation. Damage to plants may be caused by insects, mites, rodents, pets, or humans. The soil may be poorly drained, gravelly, or overly sandy; it may be covering buried debris—boards, cement blocks, bricks, and mortar; or it may be too dry or otherwise unfavourable for good plant growth. Problems also are caused by high winds, hail, lightning, blowing sand, a heavy load of snow or ice, flooding, fire, ice-removal chemicals, mechanical injury by garden tools or machinery, and fumes from weed-killing chemicals, trash burners, nearby industrial plants, or motor vehicles. The affected plant may have received treatment different from nearby healthy ones—watering, fertilizing, pest control, pruning, or depth of planting are examples. If different species or kinds of plants in the same area have similar symptoms, the chances are that a pathogen is not involved. Most infectious diseases are highly specific for individual or closely related plant species, and similar symptoms on unrelated plants are usually an indication of some environmental factor rather than a disease-causing organism.

Examination of leaves is usually considered to be the best starting point in diagnosis. The colour, size, shape, and margins of spots and blights (lesions) are often associated with a particular fungus or bacterium. Many fungi produce "signs" of disease, such as mold growth or fruiting bodies that appear as dark specks in the dead area. Early stages of bacterial infections that develop on leaves or fruits during humid weather often appear as dark and water-soaked spots with a distinct margin and sometimes a halo—a lighter-coloured ring around the spot.

Low winter temperatures and late spring or early fall freezes cause blasting (sudden death) of leaf and flower buds or sudden blighting (discoloration and death) of tender foliage.

Insect-injured leaves usually show evidence of feeding, such as holes, discoloration, stippling, blotching, downward curling, or other deformations.

Scorching of leaf margins and between the veins is common following hot, dry, windy weather. Similar symptoms are produced by an excess of water, an imbalance of essential nutrients, an excess of soluble salts, changes in the soil water table or soil grade, gas or fume injury, and root injury or disease.

Viral diseases, such as mosaics and yellows, are sometimes confused with injury by a hormone-type weed-killer, unbalanced nutrition, and soil that is excessively alkaline or acid. Nearby plant species are often examined to see if similar symptoms are evident on several different types of plants.

Diagnosis of stem system abnormalities

Examination of stems, shoots, branches, and trunk follows a thorough leaf examination. Sunken, swollen, or discoloured areas in the fleshy stem or bark may indicate canker infection by a fungus or bacterium or injury caused by excessively high or low temperatures, hail, tools, equipment, vehicles, or girdling wires.

Fruiting bodies of fungi in or on such areas often indicate secondary infection. Accurate identification of signs as belonging to a pathogenic organism or a secondary or saprophytic one is difficult. Tissues directly infected by pathogenic fungi or bacteria normally show a gradual change in colour or consistency. Injuries, in comparison, are usually well defined with an abrupt change from healthy to affected tissue.

Holes and sawdustlike debris are evidence of boring insects that usually invade woody plants in a low state of vigour. Other borer indications include wilting and dieback (progressive death of shoots that begins at tip and works downward). These symptoms also are produced by fungi and bacteria that invade water- and food-conducting vascular tissue.

Symptoms of wilt-inducing microorganisms include dark streaks in sapwood of wilted branches when the wood is cut through at an angle.

Abnormal suckers or water sprouts on trees can indicate careless pruning, extremes in temperature or water supply, structural injury, or disease.

Galls, which are unsightly overgrowths on stem, branch, or trunk, may indicate crown gall, insect injury, water imbalance between plant and soil, or other factors. Crown

Table 12: Plant Disease Symptoms

	description and causes	examples
Prenecrotic	symptom expression that precedes the death of cells or the disintegration of tissues	
Water-soaking	a water-soaked, translucent condition of tissues caused by water moving from host cells into intercellular spaces	late blight lesions on potato and tomato leaves; bacterial soft rot of fleshy vegetables
Wilting	temporary or permanent drooping of leaves, shoots, or entire plants from lack of water	bacterial wilt of cucumber; *Fusarium* wilt of tomato
Abnormal coloration	yellowing, reddening, bronzing, or purpling in localized areas of leaves where chlorophyll has been destroyed; may be due to a variety of causes	cabbage and aster yellows; halo blight of beans; potassium or phosphorus deficiency
	the presence of two or more colours in leaves and flowers due to a genetic abnormality is called variegation; viral infection results in "flower breaking"	tulip mosaic
Necrotic	localized or general death of cells or disintegration of tissues	
Blast	sudden blighting or death of young buds, flowers, or young fruit; failure to produce fruit or seeds	*Botrytis* blight of peony buds; oat blast
Bleeding	flow of sap, often discoloured, from a split crotch, branch stub, or other wound; usually accompanied by an odour of fermentation	bleeding canker of beech, dogwood, and maple
Blight	sudden or total discoloration and killing of large numbers of blossoms, leaves, shoots, or limbs or the entire plant; usually young tissues are attacked; the disease name is often coupled with the name of the host and the part attacked—blossom blight, twig blight, tip blight	fire blight of pome fruits; *Diplodia* or *Sphaeropsis* tip blight of conifers
Canker	a definite, dead, often sunken or swollen and cracked area on a stem, limb, trunk, tuber, or root surrounded by living tissues	anthracnose of sycamore and brambles; *Nectria* canker of hardwoods; fire blight of pome fruits
Damping-off	decay of seed in soil, rapid death of germinating seedlings before emergence, or emerged seedlings suddenly wilting, toppling over, and dying from rot at or near the soil line	preemergence damping-off and postemergence damping-off; both are common in seedbeds
Dieback	progressive browning and death of shoots, branches, and roots starting at the tips	winter injury; wet soil; excess soil nutrients; girdling cankers; stem or root rots; nematodes
Firing	drying and dying of leaves	nitrogen or potassium deficiency in corn; *Verticillium* wilt of eggplant
Fleck	a small, white to translucent spot or lesion visible through a leaf	ozone injury to many plants; necrotic fleck of lily
Mummification	final stage in certain fruit rots, in which the dried, shriveled, and wrinkled fruit is called a "mummy"	brown rot of stone fruits; black rot of apple
Net necrosis	an irregular crisscrossing of dark brown to black lines giving a netted appearance	in potato tubers of plants with virus leaf roll
Pitting	small dead areas within fleshy or woody tissue that appears healthy externally; definite sunken grooves or pits are formed	virus stem-pitting in apple and peach trunks; stony pit of pear fruit
Rot	decomposition and putrefaction of cells, later of tissues and organs; the rot may be dry, firm, watery, or mushy and characterized by such names as hard rot, soft rot, dry rot, black rot, and white rot	bacterial soft rot; berry rot; bud rot; bulb rot
Scald	blanching of young fruit, foliage, and shoot tissue; generally superficial	sunscald; apple and pear scald
Scorch	sudden death and "burning" of large, indefinite areas in leaves and fruit	toxicity from pesticides and air pollutants; drought; wind; lack or excess of some nutrient
Shot hole	dead spotting of leaves with diseased tissue dropping out, leaving small holes	bacterial spot; *Coryneum* blight of peach
Spot	a definite, localized, round to regular lesion, often with a border of a different colour, characterized as to location (leaf spot, fruit spot) and colour (brown spot, black spot); if numerous or if spots enlarge and merge, a large irregular blotch or blight may develop	gray leaf spot of tomato; black spot of rose; tar spot of maple
Staghead	an advanced form of dieback applied to a tree in which large branches in the upper crown are killed	oak wilt on bur oak; dwarf mistletoe on Douglas fir; *Armillaria* root rot of oak
Streak	narrow, elongated, somewhat superficial necrotic lesions, with irregular margins, on stems or leaf veins	virus streak of pea, raspberry, and tomato; Stewart's wilt of sweet corn
Stripe	narrow, elongated, parallel, necrotic lesions especially in leaf diseases of cereals and grasses	*Helminthosporium* stripe of barley; *Scolecotrichum* brown stripe of forage grasses
Hypoplastic	the underdevelopment of plant cells, tissues, or organs	
Abortion	halting development of an organ after partial differentiation	ergot of rye and other grasses
Chlorosis	yellowing or whitening of normal green tissue due to partial or complete failure of chlorophyll to develop	strawberry and aster yellows; genetic variegation in corn; iron deficiency of azalea
Stunting or dwarfing	the underdevelopment of the plant or some of its organs	dahlia stunt or mosaic; curly top of beans; little-leaf disease of pines
Rosetting	shortening of internodes of shoots and branches, producing a bunchy growth habit	peach and lily rosette
Hyperplastic or hypertrophic	an overdevelopment or overgrowth of plant cells, tissues, or organs; hyperplastic has come to mean an increase in number of cells, hypertrophic an increase in cell size	
Abscission or cast	early dropping of leaves, flowers, or small fruits; usually associated with premature formation of an abscission (separation) cell layer	black spot of rose; early blight of tomato; apple scab
Callus	overgrowth of tissues, often at margins of a canker or wound	*Nectria* canker of hardwoods; stem pitting of peach
Curl	distortion and crinkling of leaves or shoots resulting from unequal cell growth of opposite sides or in certain tissues	tobacco and tomato mosaic; leaf roll of potato; peach leaf curl
Epinasty	downward or outward curling and bending of a leaf or petiole	2,4-D injury to broadleaf plants; *Fusarium* wilt of tomato
Fasciation, or witches'-broom	a distortion that results in a dense, bushy overgrowth of thin, flattened, and sometimes curved shoots, flowers, fruit, and roots at a common point; usually due to adventitious (abnormally located) development of organs	witches'-broom of hackberry; hairy root of apple; leaf gall or fasciation of geranium (see also *Rosetting* under *Hypoplastic* in this table)
Metamorphosis or transformation	development of more or less normal tissues or organs in an abnormal location	crazy-top of corn and sorghum; formation of aerial potato tubers
Proliferation	continued development of an organ after it would normally stop growing	adventitious shoots in China aster and chrysanthemum from aster yellows mycoplasma
Russeting	usually a brownish, superficial roughening or corking of the epidermis of leaves, fruit, tubers, or other organs; often due to suberization (cork development) of cells following injury	spray or weather injury to apples; sweet potato scurf
Scab	roughened to crustlike, more or less circular, slightly raised or sunken lesions on the surface of leaves, stems, fruit, or tubers	apple, peach, and cucumber scab; common scab of potato
Gall, knot, or tumefaction	formation of local, fleshy to woody outgrowths or swellings; the outgrowth is often composed of unorganized cells	crown gall; black knot of plum; *Fusiform* gall rust of pine; nematode galls

gall is infectious and develops as rough, roundish galls at wounds, resulting from grafting, pruning, or cultivating.

Wood-decay fungi also enter unprotected wounds, resulting in discoloured, water-soaked, spongy, stringy, crumbly, or hard rots of living and dead wood. External evidence of wood-decay fungi are clusters of mushrooms (or toadstools) and hoof- or shelf-shaped fungal fruiting structures, called conks, punks, or brackets.

Aboveground symptoms of many root problems look alike. They include stunting of leaf and twig growth, poor foliage colour, gradual or sudden decline in vigour and productivity, shoot wilting and dieback, and even rapid death of the plant. The causes include infectious root and crown rot; nematode, insect, or rodent feeding; low temperature or lightning injury; household gas injury; poor soil type or drainage; change in soil grade; or massive removal of roots in digging utility trenches and construction.

Abnormal root growth is revealed by comparison with healthy roots. Some nematodes, such as root knot (*Meloidogyne* species), produce small to large galls in roots; other species cause affected roots to become discoloured, stubby, excessively branched, and decayed. Bacterial and fungal root rots commonly follow feeding by nematodes, insects, and rodents.

Diagnosis of a disease complex, one with two or more causes, is usually difficult and requires separation and identification of the individual causes.

Symptoms. The variety of symptoms, the internal and external expressions of disease, that result from any disease form the symptom complex, which, together with the accompanying signs, makes up the syndrome of the disease.

Classification of symptoms

Generalized symptoms may be classified as local or systemic, primary or secondary, and microscopic or macroscopic. Local symptoms are physiological or structural changes within a limited area of host tissue, such as leaf spots, galls, and cankers. Systemic symptoms are those involving the reaction of a greater part or all of the plant, such as wilting, yellowing, and dwarfing. Primary symptoms are the direct result of pathogen activity on invaded tissues (*e.g.,* swollen "clubs" in clubroot of cabbage and "galls" formed by feeding of the root-knot nematode). Secondary symptoms result from the physiological effects of disease on distant tissues and uninvaded organs (*e.g.,* wilting and drooping of cabbage leaves in hot weather resulting from clubroot or root knot). Microscopic disease symptoms are expressions of disease in cell structure or cell arrangement seen under a microscope. Macroscopic symptoms are expressions of disease that can be seen with the unaided eye. Specific macroscopic symptoms are classified under one of four major categories: prenecrotic, necrotic, hypoplastic, and hyperplastic or hypertrophic. These categories reflect abnormal effects on host cells, tissues, and organs that can be seen without a hand lens or microscope. See Table 12 for examples of the main disease symptoms that are classified in these four categories.

Signs. Besides symptoms, the diagnostician recognizes signs characteristic of specific diseases. Signs are either structures formed by the pathogen or the result of interaction between pathogen and host—*e.g.,* ooze of fire blight bacteria, slime flux from wetwood of elm, odour of tissues affected with bacterial soft rot. See Table 13 for the most frequently encountered signs of pathogen presence and examples of organisms producing them.

Technological advances in the identification of pathogenic agents. Developments in microscopy, serology and immunology, molecular biology, and laboratory instrumentation have resulted in many new and sophisticated laboratory procedures for the identification of plant pathogens, particularly bacteria, viruses, and viroids. The techniques of traditional scanning microscopy and transmission electron microscopy have been applied to immunosorbent electron microscopy, in which the specimen is subject to an antigen-antibody reaction before observation and scanning tunneling microscopy, which provides information about the surface of a specimen by constructing a three-dimensional image.

Serological tests have been made more specific and convenient to perform since the discovery of a technique to produce large quantities of monoclonal antibodies, which bind to only one specific antigen. The sensitivity of antigen-antibody detection has been significantly increased by a radioimmunoassay (RIA) procedure. In this procedure a "known" antigen is overlayed on a plastic plate to which antigen molecules adhere. A solution of antibody is applied to the same plate; if the antibody is specific to the antigen, it will combine with it. This is followed by the application of radioactively labeled antiantibody, which is allowed to react and then washed off. The radioactivity that remains on the plate is a measure of the amount of antibody that combined with the known fixed antigen. Another highly sensitive immunoassay is the enzyme-linked immunosorbent assay (ELISA). In principle this assay is similar to the RIA except that an enzyme system, instead of radioactivity, is used as an indicator of an antigen-antibody combination.

New analytic methods in molecular biology have made genetic studies for the characterization and identification of bacteria more practical. The DNA hybridization technique is an example. A strand of DNA from a known species (the probe) is radioactively labeled and "mixed" with DNA from an unidentified species. If the probe and the unknown DNA are from identical species, they will have complementary DNA sequences that enable them to bind to one another. Bound to DNA from the unknown species, the probe acts as a marker and identifies the bacteria.

The growing demand for quick identification of microorganisms has resulted in the development of instrumentation for automated technology that allows a large number of tests to be performed on many specimens in a short period of time. The results are read automatically and analyzed by a computer program to identify the pathogens.

PRINCIPLES OF DISEASE CONTROL

Successful disease control requires thorough knowledge of the causal agent and the disease cycle, host-pathogen interactions in relation to environmental factors, and cost. Disease control starts with the best variety, seed, or planting stock available and continues throughout the life of the plant. For harvested crops, disease control extends through transport, storage, and marketing. Relatively few diseases are controlled by a single method; the majority require several approaches. These often need to be integrated into a broad program of biological, cultural, and chemical methods to control as many different pests—including insects, mites, rodents, and weeds—on a given crop as possible.

Most control measures are directed against inoculum of the pathogen and involve the principles of exclusion and avoidance, eradication, protection, host resistance and selection, and therapy.

Exclusion and avoidance. The principle of exclusion and avoidance is to keep the pathogen away from the growing host plant. This practice commonly excludes pathogens by disinfection of plants, seeds, or other parts, using chemicals or heat. Inspection and certification of seed and other planting stock help ensure freedom from disease. For gardeners this involves sorting bulbs or corms before planting and rejecting diseased plants. Federal and state plant quarantines, or embargoes, have been established to prevent introduction of potentially destructive pathogens into areas currently free of the disease. More than 150 countries now have established quarantine regulations.

Eradication. Eradication is concerned with elimination of the disease agent after it has become established in the area of the growing host or has penetrated the host. Such measures include crop rotation, destruction of the diseased plants, elimination of alternate host plants, pruning, disinfection, and heat treatments.

Crop rotation with nonsusceptible crops "starves out" bacteria, fungi, and nematodes with a restricted host range. Some pathogens can survive only as long as the host residue persists, usually no more than a year or two. Many pathogens, however, are relatively unaffected by rotation because they become established as saprophytes in the soil (*e.g., Fusarium* and *Pythium* species; *Rhizoctonia solani;* and the potato scab actinomycete, *Streptomyces scabies*) or their propagative structures remain dormant but viable

Table 13: Signs of Pathogen Presence in Diseased Plants*

sign	description	examples
Acervulus	a shallow, saucer-shaped fungal structure that bears asexual spores (conidia); it is usually formed below the cuticle or epidermis of leaves, stems, and fruits, later rupturing the surface and exposing its spore-bearing surface	anthracnose of muskmelon and tomato; *Marssonina* leaf spot and twig blight of poplar
Apothecium	a disk-, saucer-, or cup-shaped fungal structure that produces sexual spores (ascospores); it is often stalked and fleshy	brown rot of stone fruits; *Sclerotinia* white mold of fleshy vegetables
Cleistothecium	a speck-sized, black fruiting body completely enclosing sexual spores	many powdery mildew fungi
Conidiophores	asexual fungal structures of various colours that bear conidia and appear powdery, velvety, or downy en masse; they often cover lesions of leaf, stem, or fruit	*Botrytis* blight or gray mold of many flowers; *Penicillium* mold of citrus fruit; downy mildew of grape
Conk or punk	fruiting body (sporophore) of wood-rotting fungi that produces tremendous numbers of spores (up to 100 billion per day); conks are usually large and woody and are found on tree stumps, branches, or trunks	*Fomes* and *Polyporus* wood rots of hardwoods and conifers
Mushrooms (toadstools)	fleshy, umbrella-shaped fruiting bodies of wood-decay fungi	*Armillaria* and *Clitocybe* root rots
Mycelium	the vegetative body of a fungus, which is composed of a mass of branched filaments (hyphae) often interwoven into a feltlike or woolly mass	*Rhizopus* soft rot of sweet potato and leak of strawberry; *Sclerotinia* white mold of beans
Nematode cysts	round to lemon-shaped, speck-sized bodies, white to brown in colour, are diagnostic for cyst nematodes; they are often evident on the root surface	sugar beet, soybean, and clover cyst nematodes
Odours	the process of host colonization and many pathogens give off characteristic odours	bacterial soft rot; stinking smut or bunt of wheat; slime flux of elm
Ooze or exudate	droplets of bacteria or fungal spores, usually mixed with host cell decomposition products, found on surfaces of lesions	ooze from fire blight; scab on cucumber fruit; cut stem of cucumber affected with bacterial wilt
Perithecium	speck-sized fungal fruiting body that produces large numbers of sexual spores; perithecia are dark-coloured, round to flask-shaped, usually partially buried in diseased tissue; they resemble pycnidia	apple and pear scab; *Gibberella* stalk and ear rot of corn
Powdery mildew	white, powdery to mealy, superficial growths of mycelia and conidiophores on surfaces of leaves, stems, flowers, and fruit	powdery mildew diseases of bluegrass, phlox, zinnia, and rose (see also *Cleistothecium,* this table)
Pycnidium	speck-sized fungal fruiting body that produces large numbers of asexual spores (conidia); pycnidia are dark-coloured, round to flask-shaped, usually partially buried in diseased tissue; they resemble perithecia	*Septoria* leaf spots; *Diplodia* stalk rot of corn
Rhizomorphs	cordlike or rootlike strands, composed of a bundle of closely interwined hyphae, by which certain fungi make their way through soil and over or under bark of woody plants	*Armillaria* and *Clitocybe* root rots; *Sclerotium rolfsii* stem rot of peanuts
Sclerotium	brown to black, compact, hard resting body of certain fungi with a rindlike covering; the size varies from a fly speck to a large sweet potato depending on the fungus forming it	ergot of rye; onion white rot; *Verticillium albo-atrum*
Seed	dodder seed is a sign of this parasitic flowering plant when found in clover or alfalfa seed	dodder (*Cuscuta,* about 170 species)
Sorus (pustule)	a compact mass of spores, or a cluster of sporangia (spore-bearing structures), produced in or on the host by fungi causing such diseases as white rust, smut, and true rust; before rupturing, the sorus is normally covered by host epidermis	white rust of crucifers; corn and bluegrass smuts; black stem rust of cereals
Spores	microscopic, usually single- or few-celled reproductive bodies of fungi corresponding in function to seeds of higher plants; spores vary greatly in size, shape, and colour; they are asexually produced or result from sexual processes; asexual spores may be formed directly from vegetative hyphae but often are produced in special fruiting structures (*e.g.,* acervulus, coremium, pycnidium, and sporodochium)	
Sporodochium	a cushion-shaped stroma covered with conidiophores bearing asexual spores; found scattered in leaf, stem, and fruit lesions	*Cercospora* leaf spot of celery and sugar beet; brown rot of stone fruits; *Fusarium* blight of bluegrass
Stroma	a crustlike or cushionlike mass of fungal hyphae often intermingled with host tissue on or in which spores are produced—usually in reproductive bodies	tar spot of maple and sycamore
Synnema or coremium	a tight cluster of erect conidiophores forming an elongated column on which asexual spores are borne	Dutch elm disease; oak wilt; black rot of sweet potato

*The structures listed are formed by the pathogen.

for many years (*e.g.,* cysts of cyst nematodes, sporangia of the cabbage clubroot fungus, and onion smut spores).

Burning and destruction

Burning, deep plowing of plant debris, and fall spraying are used against such diseases as leaf blights of tomato, Dutch elm disease, and apple scab. Destruction of weed hosts also helps control such viral diseases as cucumber mosaic and curly top. For fungi whose complete life cycle requires two different host species, such as black stem rust of cereals and white-pine blister rust, destruction of alternate hosts is effective. Destruction of diseased plants helps control Dutch elm disease, oak wilt, and peach viral diseases—mosaic, phony peach, and rosette. Elimination of citrus canker in the southeastern United States has been one of the few successful eradication programs in history. Infected trees were sprayed with oil and burned.

Pruning and excision of a diseased portion of the plant have aided in reducing inoculum sources for canker and wood-rot diseases of shade trees and fire blight of pome fruits. Disinfection of contaminated tools, as well as packing and shipping containers, controls a wide range of diseases. Direct application of dry or wet heat is used to obtain seeds, bulbs, other propagative materials, and even entire plants free of viruses, nematodes, and other pathogens.

Protection. The principle of protection involves placing a barrier between the pathogen and the susceptible part of the host to shield the host from the pathogen. This can be accomplished by regulation of the environment, cultural and handling practices, control of insect carriers, and application of chemical pesticides.

Regulation of the environment. Selection of outdoor growing areas where weather is unfavourable for disease is a method of controlling disease by regulating the environment. Control of viral diseases of potato, for example, can be accomplished by growing the seed crop in northern regions where low temperatures are unfavourable for the aphid carriers. Another environmental factor that can be brought under control is the storage and in-transit environment. A variety of postharvest diseases of potato, sweet potato, onion, cabbage, apple, pear, and other crops are controlled in storage and shipment by keeping humidity and temperature low and by reducing the quantity of ethylene and other natural gases in storage houses.

Cultural practices. Selection of the best time and depth of seeding and planting is an effective cultural practice that reduces disease impact. Shallow planting of potatoes may help to prevent *Rhizoctonia* canker. Early fall seeding of winter wheat may be unfavourable for seedling infection by wheat-bunt teliospores. Cool-temperature crops can be grown in soils infested with root-knot nematode and harvested before soil temperatures become favourable for nematode activity. Adjustment of soil moisture is another

Temperature, moisture, and soil pH

cultural practice of widespread usefulness. For example, seed decay, damping-off (the destruction of seedlings at the soil line), and other seedling diseases are favoured by excessively wet soils. The presence of drain tiles in poorly drained fields and the use of ridges or beds for plants are often beneficial. Adjustment of soil pH also leads to control of some diseases. Common potato scab can be controlled by adjusting the pH to 5.2 or below; other acid-tolerant plants then must be used in crop rotation, however.

Regulation of fertility level and nutrient balance. Potash and nitrogen, and the balance between the two, may affect the incidence of certain bacterial, fungal, and viral diseases of corn, cotton, tobacco, and sugar beet. A number of microelements, including boron, iron, zinc, manganese, magnesium, copper, sulfur, and molybdenum, may cause noninfectious diseases of many crop and ornamental plants. Adjusting the soil pH, adding chelated (bound or enclosed in large organic molecules) or soluble salts to the soil, or spraying the foliage with these or similar salts is a corrective measure.

Handling practices. Late blight on potato tubers can be controlled by delaying harvest until the foliage has been killed by frost, chemicals, or mechanical beaters. Avoidance of bruises and cuts while digging, grading, and packing potatoes, sweet potatoes, and bulb crops also reduces disease incidence.

Control of insect vectors. There are many examples in which losses by bacteria, viruses, and mycoplasma-like disease agents can be reduced by controlling aphids, leafhoppers, thrips, beetles, and other carriers of these agents.

Chemical control. A variety of chemicals are available that have been designed to control plant diseases by inhibiting the growth of or by killing the disease-causing pathogens. Chemicals used to control bacteria (bactericides), fungi (fungicides), and nematodes (nematicides) may be applied to seeds, foliage, flowers, fruit, or soil. They prevent or reduce infections by utilizing various principles of disease control. Eradicants are designed to kill a pathogen that may be present in the soil, on the seeds, or on vegetative propagative organs, such as bulbs, corms, and tubers. Protectants place a chemical barrier between the plant and the pathogen. Therapeutic chemicals are applied to combat an infection in progress.

Eradicants and protectants

Soil treatments are designed to kill soil-inhabiting nematodes, fungi, and bacteria. This eradication can be accomplished using steam or chemical fumigants. Soilborne nematodes can be killed by applying granular or liquid nematicides. Most soil is treated well before planting; however, certain fungicides can be mixed with the soil at planting time.

Seeds, bulbs, corms, and tubers are frequently treated with chemicals to eradicate pathogenic bacteria, fungi, and nematodes and to protect the seeds against organisms in the soil—mainly fungi—that cause decay and damping-off. Seeds are often treated with systemic fungicides, which are absorbed and provide protection for the growing seedling.

Protective sprays and dusts applied to the foliage and fruit of crops and ornamentals include a wide range of organic chemicals designed to prevent infection. Protectants are not absorbed by or translocated through the plant; thus they protect only those parts of the plant treated before invasion by the pathogen. A second application is often necessary because the chemical may be removed by wind, rain, or irrigation or may be broken down by sunlight. New, untreated growth also is susceptible to infection. New chemicals are constantly being developed.

Biological control. Biological control of plant diseases involves the use of organisms other than humans to reduce or prevent infection by a pathogen. These organisms are called antagonists; they may occur naturally within the host's environment, or they may be purposefully applied to those parts of the potential host plant where they can act directly or indirectly on the pathogen.

Although the effects of biological control have long been observed, the mechanisms by which antagonists achieve control is not completely understood. Several methods have been observed: some antagonists produce antibiotics that kill or reduce the number of closely related pathogens; some are parasites on pathogens; and others simply compete with pathogens for available food.

Cultural practices that favour a naturally occurring antagonist and exploit its beneficial action often are effective in reducing disease. One technique is to incorporate green manure, such as alfalfa, into the soil. Saprophytic microorganisms feed on the green manure, depriving potential pathogens of available nitrogen. Another practice is to make use of suppressive soils—those in which a pathogen is known to persist but causes little damage to the crop. A likely explanation for this phenomenon is that suppressive soils harbour antagonists that compete with the pathogen for food and thereby limit the growth of the pathogen population.

Other antagonists produce substances that inhibit or kill potential pathogens occurring in close proximity. An example of this process, called antibiosis, is provided by marigold (*Tagetes* species) roots, which release terthienyls, chemicals that are toxic to several species of nematodes and fungi. Antibiosis

Only a few antagonists have been developed specifically for use in plant-disease control. Citrus trees are inoculated with an attenuated strain of tristeza virus, which effectively controls the virulent strain that causes the disease. An avirulent strain of *Agrobacterium radiobacter* (K84) can be applied to plant wounds to prevent crown gall caused by infection with *Agrobacterium tumefaciens*. Many more specific antagonists are being investigated and hold much promise for future control of disease.

Therapy. Therapeutic measures have been used much less often in plant pathology than in human or animal medicine. The recent development of systemic fungicides such as oxathiins, benzimidazoles, and pyrimidines have enabled growers to treat many plants after an infection has begun. Systemic chemicals are absorbed by and translocated within the plant, restricting the spread and development of pathogens by direct or indirect toxic effects or by increasing the ability of the host to resist infection.

Antibiotics have been developed to control various plant diseases. Most of these drugs are absorbed by and translocated throughout the plant, providing systemic therapy. Streptomycin is used against a variety of bacterial pathogens; tetracycline is able to control the growth of certain mycoplasmas; and cycloheximides offer effective control for certain diseases caused by fungi.

Host resistance and selection. Disease-resistant varieties of plants offer an effective, safe, and relatively inexpensive method of control for many crop diseases. Most available commercial varieties of crop plants bear resistance to at least one, and often several, pathogens. Resistant or immune varieties are critically important for low-value crops in which other controls are unavailable, or their expense makes them impractical. Much has been accomplished in developing disease-resistant varieties of field crops, vegetables, fruits, turf grasses, and ornamentals. Although great flexibility and potential for genetic change exist in most economically important plants, pathogens are also flexible. Sometimes, a new plant variety is developed that is highly susceptible to a previously unimportant pathogen.

Variable resistance. Resistance to disease varies among plants; it may be either total (a plant is immune to a specific pathogen) or partial (a plant is tolerant to a pathogen, suffering minimal injury). The two broad categories of resistance to plant diseases are vertical (specific) and horizontal (nonspecific). A plant variety that exhibits a high degree of resistance to a single race, or strain, of a pathogen is said to be vertically resistant; this ability usually is controlled by one or a few plant genes. Horizontal resistance, on the other hand, protects plant varieties against several strains of a pathogen, although the protection is not as complete. Horizontal resistance is more common and involves many genes. Vertical and horizontal resistance

Obtaining disease-resistant plants. Several means of obtaining disease-resistant plants are commonly employed alone or in combination. These include introduction from an outside source, selection, and induced variation. All three may be used at different stages in a continuous process; for example, varieties free from injurious insects or plant diseases may be introduced for comparison with

local varieties. The more promising lines or strains are then selected for further propagation, and they are further improved by promoting as much variation as possible through hybridization or special treatment. Finally, selection of the plants showing greatest promise takes place. Developing disease-resistant plants is a continuing process.

Special treatments for inducing gene changes include the application of mutation-inducing chemicals and irradiation with ultraviolet light and X rays. These treatments commonly induce deleterious genetic changes, but, occasionally, beneficial ones also may occur.

Methods used in breeding plants for disease resistance are similar to those used in breeding for other characters except that two organisms are involved—the host plant and the pathogen. Thus, it is necessary to know as much as possible about the nature of inheritance of the resistant characters in the host plant and the existence of physiological races or strains of the pathogen.

The use of genetic engineering in developing disease-resistant plants. The techniques of genetic engineering can be used to manipulate the genetic material of a cell in order to produce a new characteristic in an organism. Genes from plants, microbes, and animals can be recombined (recombinant DNA) and introduced into the living cells of any of these organisms.

Organisms that have had genes from other species inserted into their genome (the full complement of an organism's genes) are called transgenic. The production of pathogen-resistant transgenic plants has been achieved by this method; certain genes are inserted into the plant's genome that confer resistance to such pathogens as viruses, fungi, and insects. Transgenic plants that are tolerant to herbicides and that show improvements in other qualities also have been developed.

Apprehension about the release of transgenic plants into the environment exists, and measures to safeguard the application of this technology have been adopted. In the United States several federal agencies, such as the U.S. Department of Agriculture, the Food and Drug Administration, and the Environmental Protection Agency, regulate the use of genetically engineered organisms. From 1987 to 1994 the U.S. Department of Agriculture issued more than 1,300 permits or notifications to allow transgenic plants to be evaluated in the field. With proper regulation, this technology holds great promise for making substantial advances in the control of plant diseases.

Classification of plant diseases by causal agent

Plant diseases are often classified by their physiological effects or symptoms. Many diseases, however, produce practically identical symptoms and signs but are caused by very different microorganisms or agents, thus requiring completely different control methods. Classification according to symptoms is also inadequate because a causal agent may induce several different symptoms, even on the same plant organ, which often intergrade. Classification may be according to the species of plant affected. Host indexes (lists of diseases known to occur on certain hosts in regions, countries, or continents) are valuable in diagnosis. When an apparently new disease is found on a known host, a check into the index for the specific host often leads to identification of the causal agent. It is also possible to classify diseases according to the essential process or function that is adversely affected. The best and most widely used classification of plant diseases is based on the causal agent, such as a noninfectious agent or an infectious agent (*i.e.*, a virus, viroid, mycoplasma, bacterium, fungus, nematode, or parasitic flowering plant).

Use of the host index

NONINFECTIOUS DISEASE-CAUSING AGENTS

Noninfectious diseases, which sometimes arise very suddenly, are caused by the excess, deficiency, nonavailability, or improper balance of light, air circulation, relative humidity, water, or essential soil elements; unfavourable soil moisture-oxygen relations; extremes in soil acidity or alkalinity; high or low temperatures; pesticide injury; other poisonous chemicals in air or soil; changes in soil grade; girdling of roots; mechanical and electrical agents; and soil compaction. In addition, unfavourable preharvest and storage conditions for fruits, vegetables, and nursery stock often result in losses. The effects of noninfectious diseases can be seen on a variety of plant species growing in a given locality or environment. Many diseases and injuries caused by noninfectious agents result in heavy loss but are difficult to check or eliminate because they frequently reflect ecological factors beyond human control. Symptoms may appear several weeks or months after an environmental disturbance.

Injuries incurred from accidents, poisons, or adverse environmental disturbances often result in damaged tissues that weaken a plant, enabling bacteria, fungi, or viruses to enter and add further damage. The cause may be obvious (lightning or hail), but often it is obscure. Symptoms alone are often unreliable in identifying the causal factor. A thorough examination of recent weather patterns, the condition of surrounding plants, cultural treatments or disturbances, and soil and water tests can help reveal the nature of the disease. For examples of nonpathogenic plant diseases, see Figure 1.

Adverse environment. High temperatures may scald corn, cotton, and bean leaves and may induce formation of cankers at the soil surface of tender flax, cotton, and peanut plants. Frost injury is relatively common, but temperatures just above freezing also may cause damage, such as net necrosis (localized tissue death) in potato tubers and "silvering" of corn leaves. Isolated, thin-barked trees growing in northern climates and subjected to frequent thawing by day and freezing by night may develop dead bark cankers or vertical frost cracks on the south or southwest sides of the trunk. Alternate freezing and thawing, heaving, low air moisture, and smothering under an ice-sheet cover are damaging to alfalfa, clovers, strawberries, and grass on golf greens. Legume crowns commonly split under these conditions and are invaded by decay-forming fungi.

Low-temperature injury

The drought and dry winds that often accompany high temperatures cause stunting, wilting, blasting, marginal scorching of leaves, and dieback of shoots. Leaf scorch is common on trees in exposed locations following hot, dry, windy weather when water is lost from leaves faster than it is absorbed by roots. Leaf scorch and sudden flower drop are common indoor plant problems because the humidity in a home, an apartment, or an office is usually below 30 percent. Similar symptoms are caused by a change in soil grade, an altered water-table level, a compacted and shallow soil, paved surface over tree roots, temporary flooding or a waterlogged (oxygen-deficient) soil, girdling tree roots, salt spray near the ocean, and an injured or diseased root system. Injured plants are often very susceptible to air and soil pathogens and secondary invaders.

Blossom-end rot of tomato and pepper is prevalent when soil moisture and temperature levels fluctuate widely and calcium is low.

Poor aeration may cause blackheart in stored potatoes. Accumulation of certain gases from the respiration of apples in storage may produce apple scald and other disorders.

All plants require certain mineral elements to develop and mature in a healthy state. Macronutrients such as nitrogen, potassium, phosphorus, sulfur, calcium, and magnesium are required in substantial quantities, while micronutrients or trace elements such as boron, iron, manganese, copper, zinc, and molybdenum are needed in much smaller quantities. When the supply of any essential nutrient falls below the level required by the plant, a deficiency occurs, leading to symptoms that include stunting of plants; scorching or malformation of leaves; abnormal coloration; premature leaf, bud, and flower drop; delayed maturity or failure of flower and fruit buds to develop; and dieback of shoots.

Nutrient deficiencies

Symptoms of nutrient deficiencies vary depending on the nutrients involved, the stage of plant growth, soil moisture, and other factors; they often resemble symptoms caused by infectious agents such as bacteria or viruses.

The availability of water may affect nutrient uptake by the plant. Blossom-end rot of tomato, a disease associated with a deficiency of calcium, may occur if the water sup-

ply is irregular, even if an adequate amount of calcium is in the soil. This discontinuity in availability of water will inhibit uptake of the calcium in a quantity sufficient to nourish a fast-growing tomato plant. Necrosis at the blossom end of the fruit results. This situation generally disappears when water conditions improve.

Excess minerals can damage plants either directly, causing stunting, deformities, or dieback, or indirectly by interfering with the absorption and use of other nutrients, resulting in subsequent deficiency symptoms. A superabundance of nitrogen, for example, may cause deficiency symptoms of potassium, zinc, or other nutrient elements; a lack of or delay in flower and fruit development; and a predisposition to winter injury. If potassium is high, calcium and magnesium deficiencies may occur.

The pH of a soil has a dramatic impact on nutrient availability to plants. Most plants will grow in a soil with a pH between 4.0 and 8.0. In acidic soils some nutrients are far more available and may reach concentrations that are toxic or that inhibit absorption of other nutrients, while other minerals become chemically bound and unavailable to plants. A similar situation exists in alkaline soils, although different minerals are affected. Oats planted in alkaline soils that actually contain a sufficient amount of manganese may develop the manganese-deficiency disease gray speck. This occurs because an elevated soil pH causes manganese to react with oxygen to produce manganese dioxide, a form of the nutrient that is insoluble to plants.

An excess of water-soluble salts is a common problem with houseplants. Salt concentrations may build up as a whitish crust on soil and container surfaces of potted plants following normal evaporation of water over a period of time. Symptoms include leaf scorching, bronzing, yellowing and stunting, and wilting, plus root and shoot dieback. Damage from soluble salts is also common in arid regions and in regions where ice-control chemicals are applied heavily.

Several nonparasitic diseases (*e.g.*, oat blast, weakneck of sorghum, straighthead of rice, and crazy-top of cotton) are caused by combinations of environmental factors—*e.g.*, high temperatures, moisture stress or poor irrigation practices, imbalance of mineral nutrients, and reduced light. Environmental disturbances alter the normal physiology of the plant, activity of pathogens, and host-pathogen interactions.

Toxic chemicals. Many complex chemicals are routinely applied to plants to prevent attack by insects, mites, and pathogens; to kill weeds; or to control growth. Serious damage may result when fertilizers, herbicides, fumigants, growth regulators, antidesiccants, insecticides, miticides, fungicides, nematicides, and surfactants (substances with enhanced wetting, dispersing, or cleansing properties, such as detergents) are applied at excessive rates or under hot, cold, or slow-drying conditions.

Some pollutants are the direct products of industry and fuel combustion, while others are the result of photochemical reactions between products of combustion and naturally occurring atmospheric compounds. The major pollutants toxic to plants are sulfur dioxide, fluorine, ozone, and peroxyacetyl nitrate.

Sulfur dioxide results primarily from the burning of large amounts of soft coal and high-sulfur oil. It is toxic to a wide range of plants at concentrations as low as 0.25 part per million (ppm) of air (*i.e.*, on a volume basis, one part per million represents one volume of pure gaseous toxic substance mixed in one million volumes of air) for 8 to 24 hours. Gaseous and particulate fluorides are more toxic to sensitive plants than is sulfur dioxide because they are accumulated by leaves. They are also toxic to animals that feed on such foliage. Fluorine injury is common near metal-ore smelters, refineries, and industries making fertilizers, ceramics, aluminum, glass, and bricks.

Ozone and peroxyacetyl nitrate injury (also called oxidant injury) are more prevalent in and near cities with heavy traffic problems. Exhaust gases from internal combustion engines contain large amounts of hydrocarbons (substances that principally contain carbon and hydrogen molecules—gasoline, for example). Smaller amounts of unconsumed hydrocarbons are formed by combustion of fossil fuels (*e.g.*, coal, oil, natural gas) and refuse burning. Ozone, peroxyacetyl nitrate, and other oxidizing chemicals (smog) are formed when sunlight reacts with nitrogen oxides and hydrocarbons. This pollutant complex is damaging to susceptible plants many kilometres from its source. Ozone and peroxyacetyl nitrate are capable of causing injury if present at levels of 0.01 to 0.05 part per million for several hours.

Photochemical air pollutants



Figure 1: *Nonpathogenic plant diseases.*
(Top) Spruce trees damaged by acid rain in Karkonosze National Park, Poland. (Bottom left) Molybdenum deficiency in cauliflower. (Bottom right) Blossom-end rot of tomato caused by unbalanced moisture and calcium deficiency.
(Top) © Simon Fraser/Science Photo Library—Photo Researchers, Inc.; (bottom left and bottom right) © Nigel Cattlin/Holt Studios International—Photo Researchers, Inc.

Physical injury. Lightning, hail, high winds, ice and snow loads, machinery, insect and animal feeding, and various cultural practices may seriously injure plants or plant products. With the exception of lightning, which may cause death of trees and succulent crop plants in limited areas, such injury does not usually kill plants. Wounds are created, however, through which pathogens may enter.

INFECTIOUS DISEASE-CAUSING AGENTS

Plants are subject to infection by thousands of species from very diverse groups of organisms. Most are microscopic, but a few are macroscopic. The infectious agents, as previously mentioned, are called pathogens and can be grouped as follows: viruses and viroids, bacteria (including mycoplasmas and spiroplasmas, collectively referred to as mycoplasma-like organisms [MLOs]), fungi, nematodes, and parasitic seed plants.

Diseases caused by viruses and viroids. *General characteristics.* Viruses and viroids are the smallest of the infec-

tious agents. The structurally mature infectious particle is called a virion. Virions range in size from approximately 20 nanometres (0.0000008 inch) to 250–400 nanometres and are of various shapes (see also VIRUSES). Viroids differ from viruses in that they have no structural proteins, such as those that form the protein coat (capsid) of the virus.

Both viruses and viroids are obligate parasites—*i.e.,* they are able to multiply or replicate only within a living cell of a particular host. A single plant species may be susceptible to infection by several different viruses or viroids. Major disease of important food crops such as potato, tomato, wheat, oats, rice, corn, peach, orange, sugar beet, sugarcane, and palm result from viral infection. Diseases are generally most serious in plants that are propagated vegetatively, or asexually—*i.e.,* grown from cuttings, cut divisions, sprouts, and other plant material—rather than grown from seeds (sexually propagated). For examples of diseases caused by viruses, see Figure 2.

Symptoms. The symptoms of viral and viroid plant diseases fall into four groups: (1) change in colour—yellowing, green and yellow mottling, and vein clearing; (2) malformations—distortion of leaves and flowers, rosetting, proliferation and witches'-brooms (abnormal proliferation of shoots), and little or no leaf development between the veins; (3) necrosis—leaf spots, ring spots, streaks, wilting or drooping, and internal death, especially of phloem (food-conducting) tissue; and (4) stunting or dwarfing of leaves, stems, or entire plants. Rarely they may kill the host in a short time (*e.g.,* spotted wilt and curly top of tomato). More commonly they cause reduced yield and lower quality of product.

(Left) © Mike Slater/Oxford Scientific Films; (right) © Kathy Merrifield/Photo Researchers, Inc.



Figure 2: *Plant diseases caused by viruses.*
(Left) Leaf of *Abutilon* infected by abutilon mosaic virus. (Right) Russet potatoes with corky ring spot, which is caused by tobacco rattle virus.

In many cases, virus-infected plants are more susceptible to root rots, stem or stalk rots, seedling blights, and possibly other types of diseases.

Some plants may carry one or more viruses and show no symptoms; thus, they are latent carriers and a source of infection for other plants. Symptoms of certain virus-infected plants, such as geraniums, may be masked at high temperatures. Virus symptoms reappear when the weather cools.

Virus strains

For convenience, viral/viroid diseases are often grouped together generally by symptoms, regardless of true viral/viroid relationships. Viruses also can be grouped into strains, each differing greatly in virulence and other properties. For example, two virus strains, chemically distinct, may produce indistinguishable symptoms in one orchid plant but strikingly different symptoms in another. Diseases caused by unrelated viruses may resemble one another more closely than diseases caused by strains of the same virus. Certain variegated plants, such as *Abutilon* and Rembrandt tulips, owe their horticultural uniqueness and desirability to being inherently virus-infected.

Transmission. With the exception of tobacco mosaic virus, relatively few viruses or viroids are spread extensively in the field by contact between diseased and healthy leaves.

All viruses that spread within their host tissues (systemically) can be transmitted by grafting branches or buds from diseased plants on healthy plants. Natural grafting and transmission are possible by root grafts and with dodder (*Cuscuta* species). Vegetative propagation often spreads plant viruses. Fifty to 60 viruses are transmitted in seed, and a few seed-borne viruses, such as sour-cherry yellows, are carried in pollen and transmitted by insects.

Most disease-causing viruses are carried and transmitted naturally by insects and mites, which are called vectors of the virus. The principal virus-carrying insects are about 200 species of aphids, which transmit mostly mosaic viruses, and more than 100 species of leafhoppers, which carry yellows-type viruses. Whiteflies, thrips, mealybugs, plant hoppers, grasshoppers, scales, and a few beetles also serve as vectors for certain viruses. Some viruses may persist for weeks or months and even duplicate themselves in their insect vectors; others are carried for less than an hour. Slugs, snails, birds, rabbits, and dogs also transmit a few viruses, but this is not common.

A small number of plant viruses are soilborne. Viruses causing grape fanleaf, tobacco rattle, and tobacco and tomato ring spots, as well as several strawberry viruses, are spread by nematodes feeding externally (*i.e.,* ectoparasitic) on plant roots. A few soilborne viruses may be spread by the swimming spores of primitive, soil-inhabiting pathogenic fungi, such as those causing big vein of lettuce, soilborne wheat mosaic, and tobacco necrosis.

Viruses often overwinter in biennial and perennial crops and weeds (plants that overwinter by means of roots and produce seed in their second year or during several years, respectively), in plant debris, and in insect vectors. Plants, once infected, normally remain so for life.

Control. After a plant is infected with a virus/viroid, little can be done to restore its health. Control is accomplished by several methods, such as growing resistant species and varieties of plants or obtaining virus-free seed, cuttings, or plants as a result of indexing and certification programs. Indexing is a procedure to determine the presence or absence of viruses not readily transmitted mechanically. Material from a "test" plant is grafted to an "indicator" plant that develops characteristic symptoms if affected by the viral disease in question. In addition, more drastic measures are sometimes followed, including destroying (roguing) infected crop and weed host plants and enforcing state and national quarantines or embargoes. Further control measures include controlling insect vectors by spraying plants with contact insecticides or fumigating soil to kill insects, nematodes, and other possible vectors. Growing valuable plants under fine cheesecloth or wire screening that excludes insect vectors also is done. Separation of new from virus-infected plantings of the same or closely related species is sometimes effective, and the simple practice of not propagating from plants suspected or known to harbour a virus also reduces loss.

Temperature treatments for virus control

Infected peach, apple, and rose budwood stock and carnations have been grown for weeks or months at temperatures about 37° to 38° C (99° to 100° F) to free new growth from viruses. Soaking some woody plant parts or virus-infected sugarcane shoots in hot water at about 50° C (120° F) for short periods also is effective. Both dry and wet heat treatments are based on the sensitivity of certain viruses to high temperatures. Rapidly growing dahlia and chrysanthemum sprouts outgrow viruses so that stem tips can be used to propagate healthy plants. With certain carnations, chrysanthemums, and potatoes, a few cells from the growing tip have been grown under sterile conditions in tissue culture; from these, whole plants have been developed free from viruses.

Examples of virus and viroid diseases are characterized in Table 14.

Diseases caused by bacteria. Thousands of bacterial species occur in nature. Many of these perform biochemical processes essential for the continuity of life; for example, bacterial detritivores, or decomposers, feed on nonliving organic matter, recycling it through the ecosys-

tem. There are, however, hundreds of bacterial species that cause diseases in humans, animals, and plants. For examples of diseases caused by bacteria, see Figure 3.

General characteristics. Bacteria are prokaryotic microorganisms—*i.e.*, single-celled microorganisms in which the nuclear substance is not enclosed in a membrane (see also BACTERIA). There are two major types of bacteria, the eubacteria and the archaebacteria, and they are distinguished by differences in the composition of their cell wall and cytoplasmic membrane and by certain metabolic features. Plant pathogens belong to the eubacteria. The eubacteria can be divided into three groups: gram-negative bacteria, gram-positive bacteria, and the mycoplasmas and spiroplasmas, referred to as mycoplasma-like organisms (MLOs). Gram-negative and gram-positive bacteria are distinguished on the basis of their cell wall structure, which affects the ability of the bacterium to react to the Gram stain—one of the most useful stains in bacteriologic laboratories. The distinguishing characteristic of MLOs is their lack of a cell wall; their outer boundary is instead a cytoplasmic membrane, which imparts some unusual properties not found in most eubacteria. MLOs belong to the taxonomic class Mollicutes. Plant diseases caused by MLOs are grouped as agents of "decline" (characterized by loss of vigour, decrease in yield of fruit, and eventual death) and agents of virescence (the greening of flowers) and developmental abnormalities.

The principal genera of plant pathogenic bacteria are *Agrobacterium, Clavibacter, Erwinia, Pseudomonas, Xanthomonas, Streptomyces,* and *Xylella.* With the exception of *Streptomyces* species, all are small, single, rod-shaped cells approximately 0.5 to 1.0 micrometre (0.00002 to 0.00004 inch) in width and 1.0 to 3.5 micrometres in length. Streptomycetes develop branched mycelia (narrow, threadlike growth) with curled chains of conidia (spores) on the tips of the mycelia. *Streptomyces* are gram-positive; most species of the other genera are gram-negative.

Symptoms and signs. Bacterial diseases can be grouped into four broad categories based on the extent of damage to plant tissue and the symptoms that they cause, which may include vascular wilt, necrosis, soft rot, and tumours. Vascular wilt results from the bacterial invasion of the plant's vascular system. The subsequent multiplication and blockage prevents movement (translocation) of water and nutrients through the xylem of the host plant. Drooping, wilting, or death of the aerial plant structure may occur; examples include bacterial wilt of sweet corn, alfalfa, tobacco, tomato, and cucurbits (*e.g.*, squash, pumpkin, and cucumber) and black rot of crucifers. Pathogens can cause necrosis by secreting a toxin (poison). Symptoms include formation of leaf spots, stem blights, or cankers. Soft rot diseases are caused by pathogens that secrete enzymes capable of decomposing cell wall structures, thereby destroying the texture of plant tissue—*i.e.*, the plant tissue becomes macerated (soft and watery). Soft rots commonly occur on fleshy vegetables such as potato, carrot, eggplant, squash, and tomato. Tumour diseases are caused



Figure 3: *Diseases caused by bacteria.*
(Top Left) Crown gall on euonymus. (Top right) Common scab of potato. (Bottom) Bacterial bean blight (*Xanthomonas phaseoli*).
(Top left) Runk/Schoenberger from Grant Heilman; (top right) © Nigel Cattlin/Holt Studios International—Photo Researchers, Inc.; (bottom) courtesy of Dr. James G. Kantzes, University of Maryland, Dept. of Botany

Table 14: Some Viral and Viroid Diseases of Plants

disease	causative agent	hosts	symptoms and signs	additional features
Tobacco mosaic	tobacco mosaic virus (TMV)	tobacco, tomato, and hundreds of other vegetables and weeds	mottled appearance of leaves (mosaic pattern); dwarfing	virus remains viable for years in soil and tobacco; the disease occurs worldwide; significant economic losses can occur
Cucumber mosaic	cucumber mosaic virus (CMV)	cucumber, bean, tobacco, and other plants (wide range of hosts)	similar to those of TMV infections	worldwide occurrence; very broad range of hosts
Barley yellow dwarf	barley yellow dwarf virus (BYDV)	barley, oats, rye, wheat; also pasture grasses and weeds	yellowing and dwarfing of leaves; stunting of plants	one of the most important diseases of small grains
Tomato spotted wilt	tomato spotted wilt virus (TSWV)	tomato, pepper, pineapple, peanut, and many other plants	leaves show concentric, necrotic rings; necrotic region yellow, then turning red-brown	very wide host range; infects hundreds of different plants
Prunus necrotic ring spot	prunus necrotic ring spot virus (PNRV)	stone fruits—*e.g.*, cherry, almond, peach, apricot, plum, and others	delayed foliation; leaves on infected branches show light green spots and dark rings, then become necrotic and fall off	very widespread disease of stone fruits; affects almost all trees in fruit-producing regions
Potato spindle tuber	potato spindle tuber viroid (PSTV)	potato and tomato	stunted growth; tubers are spindle-shaped and smaller than healthy tubers	the first identified viroid infection in plants; can cause major reduction in crop yield
Citrus exocortis	citrus exocortis viroid (CEV)	orange, lemon, lime, and other citrus plants	infected trees show vertical splits in bark, thin strips of partially loosened bark, and a cracked, scaly appearance	worldwide distribution; causes reduction of crop yield

by bacteria that stimulate uncontrolled multiplication of plant cells, resulting in the formation of abnormally large structures.

Most bacteria produce one major symptom; a few produce a range or combination of symptoms such as those shown in Table 12. In general, it is not particularly difficult to tell whether a plant is affected by a bacterial pathogen; however, identification of the causative agent at the species level requires isolation and characterization of the pathogen using numerous laboratory techniques (see above *Technological advances in the identification of pathogenic agents*).

Transmission and infection. In order for a bacterium to produce a disease in a plant, the bacterium must first invade the plant tissue and multiply. Bacterial pathogens enter plants through wounds, principally produced by adverse weather conditions, humans, tools and machinery, insects, and nematodes, or through natural openings such as stomates, lenticels, hydathodes, nectar-producing glands, and leaf scars.

Most foliage invaders are spread from plant to plant by windblown rain or dust. Humans disseminate bacteria through cultivation, grafting, pruning, and transporting diseased plant material. Animals, including insects and mites, are other common transmission agents. Some bacteria, such as the causal agent of Stewart's, or bacterial, wilt of corn (*Erwinia stewartii*), not only are spread by a flea beetle but also survive over winter in this insect.

Bacterial dormancy

When conditions are unfavourable for growth and multiplication, bacteria remain dormant on or inside plant tissue. Some, such as the crown gall bacterium, may survive for months or years in the soil.

Bacterial diseases are influenced greatly by temperature and moisture. Often, a difference of only a few degrees in temperature determines whether a bacterial disease will develop. In most cases, moisture as a water film on plant surfaces is essential for establishing an infection.

Control. In general, the diseases caused by bacteria are relatively difficult to control. This is partly attributable to the speed of invasion as bacteria enter natural openings or wounds directly. Direct introduction also enables them to escape the toxic effects of chemical protectants. Losses from bacterial diseases are reduced by the use of pathogen-free seed grown in arid regions. Examples of diseases controlled by this method include bacterial blights of beans and peas, black rot of crucifers, and bacterial spot and canker of tomato. Seed treatment with hot water at about 50° C (120° F) is also effective for crucifers, cucurbits, carrot, eggplant, pepper, and tomato. Bactericidal seed compounds control some bacterial diseases, such as angular leaf spot of cotton, gladiolus scab, and soft rot of ornamentals. Rotation with nonhost crops reduces losses caused by wilt of alfalfa, blights of beans and peas, black rot of crucifers, crown gall, and bacterial spot and canker of tomato. Eradication and exclusion of host plants has been useful against citrus canker, angular leaf spot of cotton, fire blight, and crown gall. Resistant varieties of crop plants have been developed to reduce losses from wilts of alfalfa, corn, and tobacco; angular leaf spot of cotton and tobacco; and bacterial pustule of soybeans, among others. Protective insecticidal sprays help control bacterial diseases, such as wilts of sweet corn and cucurbits and soft rot of iris. Protective bactericidal sprays, paints, or drenches containing copper or antibiotics are used against bacterial blights of beans and celery, fire blight, crown gall, blackleg of delphinium, and filbert and walnut blights. Finally, sanitary measures—*i.e.,* clean plow down of crop refuse, destruction of volunteer plants and weeds, sterilization of pruning and grafting tools—as well as refraining from cultivating when foliage is wet, overhead watering and spraying of indoor plants, and late cutting or grazing of alfalfa and other crops, are useful in reducing the incidence of bacterial diseases.

The characteristics of several plant diseases caused by bacteria are summarized in Table 15.

Diseases caused by fungi. Fungi cause the great majority, an estimated two-thirds, of infectious plant diseases. They include all white and true rusts, smuts, needle casts, leaf curls, mildew, sooty molds, and anthracnoses; most leaf, fruit, and flower spots; cankers; blights; scabs, root, stem, fruit, and wood rots; wilts; leaf, shoot, and bud galls; and many others. For examples of diseases caused by fungi, see Figure 4. All economically important plants apparently are attacked by one or more fungi; often many different fungi may cause disease in one plant species.

General characteristics. The fungi represent an extremely large and diverse group of eukaryotic microorganisms (see also FUNGI). The cells, which contain a membrane-bound nucleus, are devoid of chlorophyll and have rigid cell walls. Fungi have a plantlike vegetative body consisting of microscopic branching threadlike filaments of various lengths, called hyphae (singular hypha), some of which extend into the air while others penetrate the substrate on which they grow. The hyphae are arranged into a network called a mycelium. It is the mass of the mycelium that gives fungal growth its characteristic "cot-

Table 15: Some Bacterial Diseases of Plants

disease	causative agent	hosts	symptoms and signs	additional features
Granville wilt	*Pseudomonas solanacearum*	tobacco, tomato, potato, eggplant, pepper, and other plants	stunting, yellowing, and wilting of parts above ground; roots decay and become black or brown	occurs in most countries in temperate and semitropical zones; causes crop losses of hundreds of millions of dollars
Fire blight	*Erwinia amylovora*	apple and pear	blossoms appear water-soaked and shrivel, spreads to leaves and stems, causing rapid dieback	first plant disease proved to be caused by a bacterium
Wildfire of tobacco	*Pseudomonas syringae*	tobacco	yellowish green spots on leaves	wildfire of tobacco occurs worldwide—causes losses in seedlings and field plants
Blight of beans	*Xanthomonas campestris*	beans (common blight)	yellowish green spots on leaves	most phytopathogenic xanthomonads and pseudomonads cause necrotic spots on green parts of susceptible hosts; may be localized or systemic
	Pseudomonas syringae	beans (brown spot)	small water-soaked spots on lower side of leaves enlarge, coalesce, and become necrotic	occurs worldwide; causes major economic losses
Soft rot	*Erwinia carotovora*	many fleshy-tissue fruits—*e.g.,* cabbage, carrot, celery, onion	soft decay of fleshy tissues that become mushy and soft	
Crown gall	*Agrobacterium tumefaciens*	more than 100 genera of woody and herbaceous plants	initially a small enlargement of stems or roots usually at or near the soil line, increasing in size, becoming wrinkled, and turning brown to black	the conversion of a normal cell to one that produces excessive cell multiplication is caused by a plasmid (a small circular piece of DNA) carried by the pathogenic bacterium
Aster yellows	Mycoplasma-like organism (MLO)	many vegetables, ornamentals, and weeds	chlorosis; dwarfing malformations	greatest losses suffered by carrots; transmission by leafhoppers
Citrus stubborn disease	*Spiroplasma citri* (MLO)	citrus and stone fruits and vegetables	chlorosis, yellowing of leaves, shortened internodes, wilting	first MLO pathogen of plant disease cultured



Figure 4: *Plant diseases caused by fungi.*
(Top left) Brown rot on peach. (Bottom left) Corn smut. (Top right) Black knot of plum.
(Top left) © Kathy Merrifield/Photo Researchers, Inc.; (bottom left and above right) John Colwell from Grant Heilman

tony" or "fuzzy" appearance. Fungi reproduce by a variety of methods, both asexual and sexual. They produce many kinds of spores in very large numbers. For example, the colour of a moldy piece of bread is due to the colour of a massive number of microscopic mold spores.

Symptoms and signs. In general, a fungal infection can cause local or extensive necrosis. It can also inhibit normal growth (hypotrophy) or induce excessive abnormal growth (hypertrophy or hyperplasia) in a portion of or throughout an entire plant. Symptoms associated with necrosis include leaf spots, blight, scab, rots, damping-off, anthracnose, dieback, and canker. Symptoms associated with hyperplasia include clubroot, galls, warts, and leaf curls (see Table 12).

In some instances, the fungus infecting the plant may produce growth or structures on the plant, stems, or leaves such as masses of mycelium or aggregates of spores with a characteristic appearance. These developments are referred to as signs of infection, in contrast to symptoms, which refer specifically to the plant or plant tissue (see Table 13).

Transmission. Fungi are spread primarily by spores, which are produced in abundance. The spores can be carried and disseminated by wind currents, water (splashing and rain), soil (dust), insects, birds, and the remains of plants that once were infected. Vegetative fungal cells that exist in dead plant material also can be transmitted when they come in contact with a susceptible host. The survival of vegetative cells of plant pathogenic fungi in nature depends on climatic conditions, particularly temperature and moisture. Vegetative cells can survive temperatures from −5° to 45° C (23° to 113° F); fungal spores are considerably more resistant. The germination of spores, however, is favoured by mild temperatures and high humidity.

Control. Because many thousands of fungal species can infect a broad range of plants and because each fungal species has different characteristics, a variety of practices are available to control fungal diseases. The principal control measures include the use of disease-free seed and propagating stock, the destruction of all plant materials that may harbour pathogenic fungi, crop rotation, the development and use of resistant plant varieties, and the use of chemical and biological fungicides.

Several fungal diseases are characterized in Table 16.

Diseases caused by nematodes. Nematodes parasitic on plants are active, slender, unsegmented roundworms (also called nemas or eelworms). The great majority cannot be seen with the unaided eye, because they are very small and translucent. Practically all adult forms fall within the range of 0.25 to 2 millimetres in length. About 1,200 species cause disease in plants. Probably every form of plant life is fed upon by at least one species of nematode. They usually live in soil and attack small roots, but some species inhabit and feed in bulbs, buds, stems, leaves, or flowers. For examples of diseases caused by nematodes, see Figure 5.

Mode of nematode attack. Nematodes parasitic on plants obtain food by sucking juices from them. Feeding is accomplished through a hollow, needlelike mouthpart called a spear or stylet. The nematode pushes the stylet into plant cells and injects a liquid containing enzymes, which digest plant cell contents. The liquefied contents are then sucked back into the nematode's digestive tract through the stylet. Nematode feeding lowers natural resistance, reduces vigour and yield of plants, and affords easy entrance for wilt-producing or root rot-producing fungi or bacteria and other nematodes. Nematode-infested plants are weak and often appear to suffer from drought, excessive soil moisture, sunburn or frost, a mineral deficiency or imbalance, insect injury to roots or stems, or disease.

Symptoms of nematode injury

Common symptoms of nematode injury include stunting, loss of green colour and yellowing; dieback of twigs and shoots; slow general decline; wilting on hot, bright days; and lack of response to water and fertilizer. Feeder root systems are reduced; they may be stubby or excessively branched, often discoloured, and decayed. Winterkill of orchard trees, raspberries, strawberries, ornamentals, and

Table 16: Some Fungal Diseases of Plants

disease	causative agent	hosts	symptoms and signs	additional features
Late blight of potato	*Phytophthora infestans*	potato	water-soaked dark green to black or purplish lesions with pale green margins on lower leaves, white mildew at edge of lesions	responsible for Irish famine; caused starvation and death and mass migration of population
Chestnut blight	*Endothia parasitica*	chestnut tree	yellowish to reddish brown patches appear on bark; lesions spread quickly and girdle twigs or limbs, which die	disease accidently imported from Asia; first observed in New York in 1904 and rapidly spread across the United States, practically eliminating native American chestnuts
Dutch elm disease	*Ceratocystis ulmi*	elm tree	leaves wilt, turn dull green to yellow or brown, and drop off; branches die	the causative fungus is believed to have entered Europe from Asia during World War I and was later transported to the United States (1930) on elm burl logs imported for furniture veneer; elm bark beetles spread the pathogen in the United States
Black stem rust of wheat	*Puccinia graminis*	wheat; many grasses	on wheat, rust-coloured pustules with spores, chlorosis of surrounding tissue, followed by development of black teliospores; on barberry, chlorosis and hypertrophy of infected tissue, orange spore masses	disease occurs wherever wheat is grown; in 1935 it destroyed about 60 percent of the total hard red spring wheat crop in Minnesota and South Dakota; fungus has a complex life cycle, partly on wheat and partly on the barberry plant; eradication of the barberry plant is an important control measure
Coffee rust	*Hemileia vastatrix*	coffee	orange-yellow powdery spots on lower side of leaves; centres turn brown and leaves fall	most destructive disease of coffee; has caused devastating losses in all coffee-producing countries
White-pine blister rust	*Cronartium ribicola*	white pine tree	small, discoloured, spindle-shaped cankers surrounded by narrow band of yellow-orange bark; blisters exude secretion followed by bright orange pustules	one of the most important forest diseases in the United States; currant is the alternate host, and its eradication is an important control measure
Corn smut	*Ustilago maydis*	corn	minute galls form on young corn seedlings; on older plants, large galls are produced on the silk of ears and on tassels, leaves, and stalks	occurs wherever corn is grown; may cause serious crop damage
Loose smut	*Ustilago nuda*	barley, oats, wheat	infected heads are covered with masses of olive-green spores	worldwide occurrence; destroys kernels of the infected plant
Downy mildew	many species of the family Peronosporaceae	many types of plants: grapes, grasses, vegetables, and others	yellow irregular spots appear on upper leaf surface; downy fungus growth appears on underside; leaves die	one of the first plant diseases controlled by a fungicide—*i.e.*, Bordeaux mixture, a mixture of lime and copper sulfate used on grapes
Powdery mildew	many species of the family Erysiphaceae	many types of plants: grasses, vegetables, shrubs, and trees	spots of powdery mildew growth that enlarge to cover leaves or other plant organs	one of the most common and widely spread plant diseases
Apple scab	*Venturia inaequalis*	apple	small olive-coloured areas appear on young leaves, later turn black, and may coalesce; black circular spots appear on fruit	occurs almost everywhere apples are grown; infection reduces fruit size and quality
Black spot of rose	*Diplocarpon rosae*	rose	large circular black lesions on leaves; leaves turn yellow and fall off	classified as an anthracnose, which affects leaves, stems, and fruits of many plants
Anthracnose of grape	*Elsinae ampelina*	grape	(as above)	(as above)
Nectria canker	*Nectria galligena*	apple and pear and many hardwood forest trees	initially small circular brown areas that enlarge and become depressed with raised edges; callus tissue produced around canker	one of the most important diseases of pear, apple, and hardwood forest trees
Black knot of plum and cherry	*Plowrightia morbosum*	plum and cherry	small black knotty swellings on twigs and branches	occurs primarily in the eastern half of the United States and New Zealand
Brown rot	*Monilinia fructicola*	stone fruits	brown spots on blossoms; twigs develop small sunken brown cankers; fruit develops brown spots that spread rapidly	worldwide occurrence; can cause heavy losses both in orchards and in shipment
Soft rot	*Rhizopus* species	flowers, fruits, and vegetables with fleshy organs	tissues become soft with water-soaked appearance that often spreads rapidly, followed by development of fuzzy gray mycelium and black spores	infection develops most rapidly on ripe fruits with favourable conditions (moderate temperature and high humidity)
Fusarium wilt of tomato	*Fusarium oxysporum*	tomatoes	leaves are bent down, growth is stunted, plant dies; dark streaks appear in vascular tissue	one of the most destructive diseases of tomato; entire fields can be destroyed
Wilts of vegetables, flowers, and some trees	*Verticillium* species	cotton, potato, tomato, alfalfa, shade trees, and others	similar to fusarium wilts; develops primarily in seedlings that die shortly after infection; older plants also are attacked	worldwide distribution; the fungus infects hundreds of species of plants

other perennials is commonly associated with nematode infestations.

Root injury develops partly from the nematodes feeding on cells and partly from toxic salivary excretions of the parasite. Tissues often respond by producing either an enlargement or degeneration of cells; sometimes both occur.

Many nematodes are native and attack cultivated plants when their natural hosts are removed. Others have been introduced with seedling plants, bulbs, tubers, and particularly in soil balled around roots of infested nursery stock.

Nematodes may live part of the time free in soil around roots or in fallow gardens and fields. They tunnel inside plant tissues (endoparasites) or feed externally from the surface (ectoparasites) and may enter a plant through wounds or natural openings or by penetrating roots. All nematodes parasitic on plants require living plant tissues for reproduction. Nematodes are attracted to host roots by sensing either the heat given off by roots or the chemicals secreted by roots.

Most species require 20 to 60 days to complete a generation from egg through four larval stages to adult and back to egg. Some nematodes have only one generation a year but still produce several hundred offspring.

Soil populations and developmental rate of nematodes are affected by the length of the growing season; temperature; availability of water and nutrients; and moisture, type, texture, and structure of soil. Also important are populations of nematode-parasitic bacteria, viruses, some 50 different nematode-trapping fungi, protozoans, mites, flatworms, or other pests, and other nematodes. Toxic chemicals added to the soil or those secreted by plant roots; crop rotations and past cropping history; species, variety, age and nutrition of growing plants; and other factors are additional conditions that affect nematode populations.

Certain species live strictly in light, sandy soils; some build up high populations in muck soils; and a few seem to thrive in heavy soils. High populations and greater crop damage are much more common in light sandy soils than in heavy clay soils.

Many plant-infecting nematodes become inactive at temperatures between 5° and 15° C (41° and 59° F) and 30° and 40° C (86° and 104° F). The optimum for most is 20° to 30° C (68° to 86° F), but this varies greatly with the species, stage of development, activity, growth of the host, and other factors.

Numbers of nematodes in diseased plants

Nematodes may be found in plant tissues in large numbers. Hundreds of thousands may be present in infested roots or bulbs.

After a plant-infecting nematode has been accidentally introduced into a garden or field, several years pass before the population builds up sufficiently (*i.e.,* up to several billion or more active nematodes per hectare) to cause conspicuous symptoms in a large number of plants. This is because nematodes move very slowly through soil—rarely more than 75 centimetres a year. Nematodes are easily spread, however, by moving infested soil, plant parts, or contaminated objects—*e.g.,* tools and machinery, bags and other containers, running water, wind, clothing, shoes, animals, birds, and infested planting stock.

Root-knot nematodes

Nematode diseases. Root-knot nematodes (*Meloidogyne* species) are well known because of the conspicuous "knots," or gall-like swellings, they induce on roots. More than 2,000 kinds of higher plants are subject to their attack. Losses are often heavy, especially in warm regions with long growing seasons. Certain species, however, such as the northern root-knot nematode (*M. hapla*), are found where soil may freeze to depths of nearly a metre. Vegetables, cotton, strawberry, and orchard trees are commonly attacked. Garden plants and ornamentals frequently become infested through nursery stock.

Root-lesion nematodes (*Pratylenchus* species), cosmopolitan in distribution, are endoparasites that cause severe losses to hundreds of different crop and ornamental plants by penetrating roots and making their way through the tissues, breaking down the cells as they feed. They deposit eggs from which new colonies develop. After a root begins to decline in vigour, nematodes move into the soil in search of healthy roots. Lesions form in the root as fungi and bacteria enter damaged tissues, and root rot often occurs. Annual crops may succumb early in the season, but perennials and orchard trees may not decline for several years.



Figure 5: *Plant diseases caused by nematodes.*
(Left) Aboveground symptoms on potato plants infested by the golden nematode of potatoes and a secondary fungal infection. (Right) "Knots" on root of sugar beet caused by root-knot nematodes.
(Left and right) © Nigel Cattlin/Holt Studios International—Photo Researchers, Inc.

The golden nematode of potatoes (*Heterodera rostochiensis*) is a menace of the European potato industry. Great efforts have been made to control it. The speck-sized golden cysts that dot infested plant roots are the remains of female bodies. Each cyst may contain up to 500 eggs, which hatch in the soil over a period of up to 17 years. A chemical given off by potato and tomato roots stimulates hatching of the eggs.

A related, cyst-forming species, the sugar beet nematode (*H. schachtii*), is a pest that has restricted acreage of sugar beets in Europe, Asia, and America.

The citrus nematode (*Tylenchulus semipenetrans*) occurs wherever citrus is grown, exacting a heavy toll in fruit quality and production. Typical symptoms are a slow decline, yellowing and dying of leaves, and dieback of twigs and branches in many groves 15 years or older. Infested nursery stock has widely distributed the nematode. The burrowing nematode (*Radopholus similis*) is a serious endoparasite in tropical and subtropical areas, where it attacks citrus (causing spreading decline), banana, avocado, tomato, black pepper, abaca, and more than 200 important crops, trees, and ornamentals, causing severe losses.

Root-feeding nematodes

Many important ectoparasites feed on plant roots—dagger nematodes (*Xiphinema*), stubby-root nematodes (*Trichodorus*), spiral nematodes (*Rotylenchus* and *Helicotylenchus*), sting nematodes (*Belonolaimus*), and pin nematodes (*Paratylenchus*). Leaf, or foliar, nematodes (*Aphelenchoides* species) and bulb and stem nematodes (*Ditylenchus dipsaci*) cause severe losses in vegetable and ornamental bulb crops, clovers, alfalfa, strawberry, sweet potato, orchids, chrysanthemums, begonias, and ferns.

Control measures. Control measures for nematodes often include rotation with nonhost plants, growing of resistant varieties and species, use of certified, nematode-free nursery stock, and use of soil fumigants (nematicides) as preplanting or postplanting treatments. Steam or dry heat is applied to soil in confined areas, such as greenhouse benches and ground beds. Exposure to moist heat, such as steam or hot water at 50° C (120° F) for 30 minutes, is sufficient to kill most nematodes and nematode eggs. Shorter periods are needed at higher temperatures. State and federal quarantines prohibiting movement of infested soil, plants or plant parts, machinery, and other likely carriers also exist. Cultural practices to promote vigorous plant growth (*i.e.,* watering during droughts, proper application of fertilizers, clean cultivation, fall and summer fallowing, use of heavy organic mulches or cover crops, and plowing out roots of susceptible plants after harvest) are useful for specific nematodes. Asparagus, marigolds (*Tagetes* species), and *Crotalaria* species are toxic to many plant-infecting nematodes.

Parasitic seed plants. A number of flowering plants are parasites of other plants. Among the more important ones are mistletoe, dodder, and witchweed. For an example of diseases caused by parasitic plants, see Figure 6.



Figure 6: Dodder, a seed-producing parasite, entwined around blueberry.
Courtesy of Dr. W.V. Welker, weed scientist, USDA (retired)

Mistletoe. Mistletoes are semiparasitic seed plants that feed on trees and obtain water and mineral salts by sending rootlike structures (haustoria) into vascular tissue of the inner bark. There are three important types: American (*Phorodendron* species), European (*Viscum album*), and dwarf (*Arceuthobium* species). All produce sticky seeds spread by birds. American mistletoe, restricted to the Americas, is best known for its ornamental and sentimental uses at Christmastime. The leafy, bushy evergreen masses, up to one metre or more in diameter, appear on tree branches. They are most conspicuous after deciduous leaves have fallen. The European mistletoe is similar in habit and appearance to its American relative. Tree branches infected by mistletoes may become stunted or even die.

Dwarf mistletoe

Dwarf mistletoe is common on and very destructive to conifers in forests. Seedlings and young trees may be stunted, deformed, or killed. Conspicuous witches'-brooms form in the crown or spindle-shaped swellings (later cankers) in limbs and trunk. Canker and wood-rotting fungi often enter through mistletoe wounds. Dwarf mistletoes frequently escape detection because the scaly-leaved plants may be less than 2½ centimetres long; they do range to 30 to 45 centimetres, however. Dwarf mistletoes occur scattered along conifer limbs and small branches. After the mistletoe has grown internally for about a year, the branch may start to form a witches'-broom. Four to five years elapse before the yellow to brown to olive-green shoots form fruits. The sticky seeds are shot with explosive force from the fruit for horizontal distances ranging from 5 to more than 18 metres; this is one of the most remarkable methods for seed discharge among plants. Once seeds adhere to a branch, they germinate on young bark and penetrate into the host tree's vascular system. Control for mistletoes in individual trees involves removal of infected branches a foot or more beyond any evidence of the parasite before the fruits ripen.

Dodder. More than 100 species of dodder (*Cuscuta*) are widely distributed and called such names as strangleweed, devil's-hair, pull down, hell-bind, love vine, and goldthread. The leafless, yellow-orange, threadlike stems twine around a number of field and garden host plants. By extending to nearby plants, it may draw them together and downward until a tangled yellowish orange patch is formed. The infested area is usually less than three metres across the first year; it spreads more rapidly in succeeding years. Dodder is widely distributed as a contaminant with field seed; hence the losses in clover, alfalfa, and flax fields. Dodder is controlled by planting certified, properly cleaned seed and by mowing patches of dodder in the field well before the seeds form. The dried patches are sprinkled with fuel oil and burned. Careful application of selective herbicides or a soil fumigant and sowing heavily infested areas with resistant plants (*e.g.*, garden beans, soybean, corn, cowpea, pea, grasses, or small grains) are also control methods.

Witchweed. Witchweed, a small parasitic weed (*Striga asiatica*), is widely distributed in Asia, southern Africa, and the Sahel. It has been known in the coastal sandy soils of North and South Carolina since the mid-1950s but through intensive efforts has been contained. Witchweed parasitizes the roots of many hosts, including maize (corn), sorghum, sugarcane, rice, small grains, and more than 50 species in the grass and sedge families. A serious infestation may cause corn plants to be severely stunted, wilt, and turn yellow or brown, thus reducing the acre yield. *Striga* plants, which rarely exceed heights of 20 to 25 centimetres, have small, red, yellowish red, yellow, or white flowers. One plant may produce hundreds of thousands of tiny brown seeds that can remain alive in soil for years until stimulated to germinate by a secretion from a nearby host root. Witchweed robs the host of water and food, causing it to grow more slowly than normal and often to die before maturing. Control is difficult; useful measures include application of selective herbicides before seeds are produced; rotation with a resistant crop and keeping plantings free of weed grasses that may serve as hosts; and prevention of seed set by growing trap crops and then destroying them with herbicides.

Control of witchweed

(M.C.S./Ar.Kn./M.J.P./R.M.P.)

BIBLIOGRAPHY

General works. LAWRIE REZNEK, *The Nature of Disease* (1987), written for the general reader, discusses the nature of disease from several perspectives, including medical, legal, political, philosophical, and economic. DAVID O. SLAUSON, BARRY J. COOPER, and MAJA M. SUTER, *Mechanisms of Disease: A Textbook of Comparative General Pathology*, 2nd ed. (1990), written for the veterinary student but a great resource for pathologists and biomedical researchers, provides a fundamental overview of the mechanisms of diseases, often at the molecular level. MAX SAMTER (ed.), *Immunological Diseases*, 4th ed., 2 vol. (1988), covers the collagen diseases. F.M. BURNET, *The Natural History of Infectious Disease*, 3rd ed. (1962), offers a unique view of infectious disease as an ecological and evolutionary phenomenon. Books for the general reader include JUNE GOODFIELD, *Quest for the Killers* (1985), exploring efforts to conquer several epidemic diseases; ANDREW SCOTT, *Pirates of the Cell: The Story of Viruses from Molecule to Microbe*, rev. ed. (1987); and PETER RADETSKY, *The Invisible Invaders: The Story of the Emerging Age of Viruses* (1991). (W.Bu./D.G.Sc.)

Human disease. KENNETH F. KIPLE (ed.), *The Cambridge World History of Human Disease* (1993), a reference text written for advanced undergraduates and professionals in the biomedical and social sciences, surveys the medical and geographic characteristics of human diseases worldwide throughout history. JAMES B. WYNGAARDEN, LLOYD H. SMITH, JR., and J. CLAUDE BENNETT (eds.), *Cecil Textbook of Medicine*, 19th ed. (1992), considers all facets of human disease in depth from the modern point of view. *Goodman & Gilman's The Pharmacological Basis of Therapeutics*, 9th ed. by JOEL G. HARDMAN and LEE E. LIMBIRD (1996), is a comprehensive text on drugs. T.R. HARRISON, *Harrison's Principles of Internal Medicine*, 13th ed. edited by KURT J. ISSELBACHER *et al.* (1994), discusses in detail the cardinal manifestations of disease under various headings. THEODORE LIDZ, *The Person: His and Her Development Throughout the Life Cycle*, rev. ed. (1976, reissued 1983), provides an excellent insight into humans, the psychological organisms. VINAY KUMAR, RAMZI S. COTRAN, and STANLEY L. ROBBINS, *Basic Pathology*, 5th ed. (1992), clearly and succinctly presents the causes and pathogenesis of human disease with an emphasis on molecular mechanisms. MARGARET W. THOMPSON, RODERICK R. MCINNES, and HUNTINGTON F. WILLARD, *Thompson & Thompson Genetics in Medicine*, 5th ed. (1991), is a well-illustrated and clearly written text on basic genetic principles and their relation to the genesis of human disease. CHARLES R. SCRIVER *et al.* (eds.), *The Metabolic Basis of Inherited Disease*, 6th ed., 2 vol. (1989), a monumental, highly technical text, provides a comprehensive presentation of the clinical, biochemical, and genetic information concerning those diseases thought to be a consequence of genetic variation. More specific in focus and perhaps less monumental (if not less technical) than the above are ROGER N. ROSENBERG *et al.* (eds.), *The Molecular and Genetic Basis of Neurological Disease* (1993); ALDONS J. LUSIS, JEROME I. ROTTER, and ROBERT S. SPARKES (eds.), *Molecular Genetics of Coronary Artery Disease* (1992); and LINDA L. GALLO (ed.), *Cardiovascular Disease: Molecular and Cellular Mechanisms, Prevention, and Treatment* (1987), which address their particular topics on cellular and molecular levels. ROBERT C. GALLO and FLOSSIE WONG-STAAL (eds.), *Retrovirus Biology and Human Disease* (1990), written for the technically advanced reader, covers various topics in retrovirology, including historical background, epidemiology, clinical features, molecular

biology, immunology, and therapeutic approaches. ADRIANNE BENDICH and C.E. BUTTERWORTH, JR. (eds.), *Micronutrients in Health and in Disease Prevention* (1991), discusses evidence of a correlation between the intake of nonoptimal levels of dietary micronutrients and the development of chronic diseases; although written for the health-care professional, it is also valuable to anyone interested in the relationship between nutrition and health. (S.L.R./J.H.Ro./D.G.Sc.)

Diseases of animals. CALVIN W. SCHWABE, *Veterinary Medicine and Human Health,* 3rd ed. (1984), provides a comprehensive reference on medical public health. Also of interest is PAUL R. SCHNURRENBERGER, ROBERT S. SHARMAN, and GILBERT H. WISE, *Attacking Animal Diseases: Concepts and Strategies for Control and Eradication* (1987). J.F. SMITHCORS, *Evolution of the Veterinary Art: A Narrative Account to 1850* (1957), comprehensively treats veterinary medical history and the history of the knowledge of animal diseases; it is brought up to date by D.H.V. STALHEIM, *The Winning of Animal Health. 100 Years of Veterinary Medicine* (1994). LISE WILKINSON, *Animals and Disease: An Introduction to the History of Comparative Medicine* (1992), studies the interrelation of animal and human diseases. O.M. RADOSTITS, D.C. BLOOD, and C.C. GAY, *Veterinary Medicine,* 8th ed. (1994), focuses on livestock. CARLTON L. GYLES and CHARLES O. THOEN (eds.), *Pathogenesis of Bacterial Infections in Animals* (1986); JOHN W. DAVIS, LARS H. KARSTAD, and DANIEL O. TRAINER (eds.), *Infectious Diseases of Wild Mammals,* 2nd ed. (1981); and IVAL ARTHUR MERCHANT and RALPH DAVID BARNER, *An Outline of the Infectious Diseases of Domestic Animals,* 3rd ed. (1964), are both about animal diseases. GEORGE F. BODDIE, *Diagnostic Methods in Veterinary Medicine,* 6th ed. (1969); JIRO J. KANEKO (ed.), *Clinical Biochemistry of Domestic Animals,* 4th ed. (1989); G.R. CARTER and JOHN R. COLE, JR. (eds.), *Diagnostic Procedures in Veterinary Bacteriology and Mycology,* 5th ed. (1990); and CHARLES M. FRASER *et al.* (eds.), *The Merck Veterinary Manual: A Handbook of Diagnosis, Therapy, and Disease Prevention and Control for the Veterinarian,* 7th ed. (1991), are general references on physical and laboratory diagnostic techniques. (C.E.Co./Ed.)

Diseases of plants. General works include G.C. AINSWORTH, *Introduction to the History of Plant Pathology* (1981), a review of the developments in the field of plant pathology and the influence of plant diseases on history; GAIL L. SCHUMANN, *Plant Diseases: Their Biology and Social Impact* (1991), a discussion of the social and cultural influence of plant diseases; and E.C. LARGE, *The Advance of the Fungi* (1940, reissued 1962), a popular account of plant disease epidemics and how they have influenced economic and political history.

Compilations of practical information are A. JOHNSTON and C. BOOTH (eds.), *Plant Pathologist's Pocket Book,* 2nd ed. (1983), on the identification, isolation, and culture of plant pathogens; MICHAEL D. SMITH (ed.), *The Ortho Problem Solver,* 3rd ed. (1989), a handbook for indoor and outdoor plants; *Westcott's Plant Disease Handbook,* 5th ed. rev. by R. KENNETH HORST (1990), a comprehensive reference covering plants grown in the United States, for professional and amateur gardeners; LOUIS PYENSON, *Plant Health Handbook* (1981), a guide for the amateur gardener; G.R. DIXON, *Plant Pathogens and Their Control in Horticulture* (1984), with both host and microorganism/disease indexes; I.M. SMITH *et al.* (eds.), *European Handbook of Plant Diseases* (1988), a reference for professional plant pathologists and for advanced study in the field, covering economically important diseases of crops and forest trees in Europe; PASCAL P. PIRONE, *Diseases and Pests of Ornamental Plants,* 5th ed. (1978), an excellent reference for gardeners and landscape professionals; RUBERT BURLEY STREETS, *The Diagnosis of Plant Diseases: A Field and Laboratory Manual Emphasizing the Most Practical Methods for Rapid Identification* (1982); WILLIAM R. JARVIS, *Managing Diseases in Greenhouse Crops* (1992); and H. DAVID THURSTON, *Tropical Plant Diseases* (1984), a discussion of important diseases of major tropical crops. The *Compendium of Plant Diseases* is an outstanding series of well-illustrated books by experts in each field, designed to assist in the identification, prevention, and control of major plant diseases and disorders of specific crops.

College-level texts include GEORGE N. AGRIOS, *Plant Pathology,* 3rd ed. (1988), a comprehensive discussion of parasitism and pathogenicity and the biochemistry of host-pathogen relationships; J.G. MANNERS, *Principles of Plant Pathology,* 2nd ed. (1993), an investigation of the physiology and genetics of host-pathogen interactions, disease epidemiology, and control; DANIEL A. ROBERTS and CARL W. BOOTHROYD, *Fundamentals of Plant Pathology,* 2nd ed. (1984), an examination of causal agents, symptoms, and control of plant diseases; C.H. DICKINSON and J.A. LUCAS, *Plant Pathology and Plant Pathogens,* 2nd ed. (1982), an overview of plant disease with extensive coverage of host-pathogen interactions at both the cellular and subcellular levels; and GEORGE B. LUCAS, C. LEE CAMPBELL, and LEON T. LUCAS, *Introduction to Plant Diseases: Identification and Management,* 2nd ed. (1992), a survey of the causes, impact, and management of plant diseases. Specific aspects are treated in depth in the following references: KURT J. LEONARD and WILLIAM E. FRY (eds.), *Plant Disease Epidemiology,* vol. 1 (1986), covering population dynamics and management of disease-causing agents; JÜRGEN KRANZ (ed.), *Epidemics of Plant Diseases,* 2nd completely rev. ed. (1990), a presentation of the latest mathematical and statistical methods in use for analysis and modeling of plant disease epidemics; R.K.S. WOOD and G.J. JELLIS (eds.), *Plant Diseases: Infection, Damage, and Loss* (1984), a comprehensive assessment; WILLIAM F. BENNETT (ed.), *Nutrient Deficiencies & Toxicities in Crop Plants* (1993), an examination of the role of nutrients on the health of major crop plants; R.D. DURBIN (ed.), *Toxins in Plant Disease* (1981), on the role of microbial toxins in the plant disease cycle; P.G. AYRES (ed.), *Effects of Disease on the Physiology of the Growing Plant* (1981), a compilation of seminar papers; GEORGE W. BRUEHL, *Soilborne Plant Pathogens* (1987); DAVID F. FARR *et al., Fungi on Plants and Plant Products in the United States* (1989), a comprehensive discussion; MASAO GOTO, *Fundamentals of Bacterial Plant Pathology* (1992), a discussion of the morphology, taxonomy, and physiology of phytopathogenic bacteria; J.F. BRADBURY, *Guide to Plant Pathogenic Bacteria* (1986), identifying phytopathogenic bacteria and the diseases they cause; DAVID C. SIGEE, *Bacterial Plant Pathology: Cell and Molecular Aspects* (1993), including discussions of interactions with host cells, virulence factors, and genetics; R.E.F. MATTHEWS, *Plant Virology,* 3rd ed. (1991), a comprehensive text covering all aspects of plant-infecting viruses, and *Diagnosis of Plant Virus Diseases* (1993), a discussion of strategies; R.T. PLUMB and J.M. THRESH (eds.), *Plant Virus Epidemiology* (1983), a collection of works by international authorities in the field of plant virology, including several case histories of particularly important diseases; KARL MARAMOROSCH (ed.), *Plant Diseases of Viral, Viroid, Mycoplasma, and Uncertain Etiology* (1992); KARL MARAMOROSCH and S.P. RAYCHAUDHURI (eds.), *Mycoplasma Diseases of Crops* (1988), a collection of discussions regarding detection of mycoplasmas, their interactions with plants, insects, and viruses, and their control; VICTOR H. DROPKIN, *Introduction to Plant Nematology,* 2nd ed. (1989), on the classification and characterization of plant-parasitic nematodes; M. WAJID KHAN (ed.), *Nematode Interactions* (1993); KERRY F. HARRIS and KARL MARAMOROSCH (eds.), *Pathogens, Vectors, and Plant Diseases: Approaches to Control* (1982), on the identification and control of insect and nematode vectors of plant diseases; ANNE R. LESLIE and GERRIT W. CUPERUS (eds.), *Successful Implementation of Integrated Pest Management for Agricultural Crops* (1993); H. DAVID THURSTON, *Sustainable Practices for Plant Disease Management in Traditional Farming Systems* (1992), a discussion of integrated control of phytopathogenic microorganisms; RICHARD N. STRANGE, *Plant Disease Control: Towards Environmentally Acceptable Methods* (1993), an ecologically sensitive analysis of current methods of disease prevention and control; R. JAMES COOK and KENNETH F. BAKER, *The Nature and Practice of Biological Control of Plant Pathogens* (1983), a discussion of the influence of environment on the interactions among microorganisms and crop plants; ARTHUR W. ENGELHARD (ed.), *Soilborne Plant Pathogens: Management of Diseases with Macro- and Microelements* (1989), on the effects of fertilization, nutrition, and pH on diseases caused by soilborne plant pathogens; P.R. DAY and G.J. JELLIS (eds.), *Genetics and Plant Pathogenesis* (1987), covering the genetic aspects of disease and pest resistance; and M.S. WOLFE and C.E. CATEN (eds.), *Populations of Plant Pathogens: Their Dynamics and Genetics* (1987), a collection of essays with ideas and concepts relevant to the long-term development of disease-control methods based on population dynamics. (M.J.P./R.M.P.)

Religious Doctrines and Dogmas

The development of doctrines and dogmas—*i.e.*, the explications and officially acceptable versions of religious teachings—has significantly affected the traditions, institutions, and practices of the religions of the world. Doctrines and dogmas also have influenced and been influenced by the ongoing development of secular history, science, and philosophy.

This article is divided into the following sections:

General nature of doctrine and dogma 394
- Distinctions between doctrine and dogma
- Functions of doctrines and dogmas
- Development
- The relation of faith, reason, and religious insight to doctrine and dogma
- Changing conceptions

Major themes and motifs 396
- Creation 396
  - Nature and significance
  - Types of cosmogonic myths
  - Doctrines of creation
  - Skepticism regarding creation
- Eschatology 401
  - Nature and significance
  - General characteristics
  - The forms of eschatology
  - Eschatological terminology
  - Eschatology in non-Western religions
  - Eschatology in religions of the West
  - Eschatology in modern times
- Angels and demons 408
  - Nature and significance
  - Celestial and noncelestial forms
  - Types of angels and demons
  - Varieties of angels and demons in the religions of the world
- Salvation 412
  - Nature and significance
  - Basic context
  - Methods and techniques
  - Varieties of salvation in world religions
- Providence 416
  - Nature and significance
  - Basic concepts and scope
  - Critical problems
- Revelation 418
  - Nature and significance
  - Types and variations
  - Themes and functions
  - Conclusion
- Covenant 421
  - Nature and significance
  - Origin and function of convenants
  - The origin and development of biblical covenants: Judaism
  - The origin and development of the covenant in Christianity
  - Covenant in other religions
- Prophecy 425
  - Nature and significance
  - Types of prophecy
  - Prophecy in the ancient Middle East and Israel
  - Prophecy in Christianity
  - Prophecy in Islām
  - Prophecy in other religions
- Miracle 431
  - Nature and significance
  - Types and functions of miracles
  - Sources of miracles
  - Miracles in the religions of the world
  - Interpretation of miracles
- Saint 436
  - Nature and significance
  - Saints in Eastern religions
  - Saints in Western religions
  - Modes of recognition
  - Types and functions of saints

Bibliography 440

GENERAL NATURE OF DOCTRINE AND DOGMA

DISTINCTIONS BETWEEN DOCTRINE AND DOGMA

Doctrine in theology (Latin *doctrina*; Greek *didaskalia, didachē*) is a generic term for the theoretical component of religious experience. It signifies the process of conceptualizing the primal—often experiential or intuitive—insights of the faith of a religious community in support of rationally understood belief. Doctrines seek to provide religion with intellectual systems for guidance in the processes of instruction, discipline, propaganda, and controversy. Dogma (Latin *decretum*, Greek *dogma*) has come to have a more specific reference to the distillate of doctrines: those first (basic or axiomatic) principles at the heart of doctrinal reflection, professed as essential by all the faithful.

This distinction appears in Christianity in the New Testament, in which *didaskalia* means "basic teachings" (as in I and II Tim.) whereas *dogma* is used only in the sense of an official judgment or decree (as in Acts 16:4). Later, however, many theologians of the early church (including, for example, Origen, St. Cyril of Jerusalem, and St. Jerome) use the term dogma in the sense of doctrine. In Eastern Christianity, the theologian St. John of Damascus popularized the term "orthodoxy" (literally "correct views") to connote the sum of Christian truth. In Western Christianity, the great medieval theologian St. Thomas Aquinas chose the phrase "articles of faith" to denote those doctrines that are solemnly defined by the church and are considered to be obligatory for faith. As late as the Roman Catholic reformatory Council of Trent (1545–63), "doctrine" and "dogma" were still roughly synonymous.

Most modern historians, however, have stressed their difference. According to J.K.L. Gieseler, a 19th-century German church historian, in *Dogmengeschichte,*

> Dogma is not doctrinal opinion, not the pronouncement of any given teacher, but doctrinal statute (*decretum*). The dogmas of a church are those doctrines which it declares to be the most essential contents of Christianity.

A modern church historian, Adolf von Harnack, sought to explain the rise of dogma in Christianity as the specific consequence of an alien blend of Greek metaphysics and Christian thought that had been rendered obsolete by Protestantism's appeal to Scripture and history. The German Roman Catholic dogmatician Karl Rahner's contrasting definition, in *Sacramentum Mundi,* points to a perennial process:

> Dogma is a *form* of the abiding vitality of the deposit of faith in the church which itself remains always the same.

FUNCTIONS OF DOCTRINES AND DOGMAS

The functions of doctrines and dogmas vary in the several religious traditions according to the stress each puts on the importance of the rational conceptualization of religious truth first glimpsed in images, symbols, and parables. In what are viewed by some scholars as the more mystical religions of the East, doctrines are usually designed to serve as catalytic clues to religious insight (*e.g.*, the notions of Nirvāna, or the goal of the religious life, in Hinduism, Jainism, and Buddhism). In what are regarded as the more

The rational conceptualization of religious truth

personalistic religions of the West, doctrines and dogmas tend to function as aids to theological reflection (*e.g.,* the concept of God's unity in Judaism, Christianity, and Islām). In all the higher religions, doctrines and dogmas emerge and develop in the service of instruction for the faithful: interpreting their sacred Scriptures, understanding their obligations and duties, and safeguarding the lines between allowable diversity and actual error—all of which help to chart the religious pathway to wisdom, rectitude, and fulfillment. Theology (which utilizes doctrines and dogmas) is, according to the medieval Christian theologian and churchman St. Anselm of Canterbury, "faith seeking rational self-understanding."

The normative function of doctrinal formulation is a typically vain effort to fix and conserve an interpretation of the original dogmas of a given tradition. The themes of *saṃsāra* (the process of reincarnation) and *karman* (the law of cause and effect) are shared by Hinduism, Jainism, and Buddhism, though with quite different doctrinal explicitations and consequences. Analogous developments are evident in other traditions.

A third function of doctrine is polemical: the defense of the faith against misinterpretation and error, within or without a religious tradition. Given the invariably pluralistic character of theological reflection, there is a constant tension between the concern for identity and continuity of the tradition, on the one hand, and for deeper and richer comprehension of truth itself, on the other. Over against this there is in most cultures a concurrent rivalry with other religions, with their contrary doctrinal claims, and beyond that, the challenges of secular wisdom and unbelief. This calls forth a special sort of doctrinal formulation: apologetics, the vindication of the true faith against its detractors or disbelievers.

The problem of primal authority

At the heart of all efforts to support religious faith lies the problem of primal authority. It is required of a doctrinal statement that it be clear and cogent, but doctrines always point past their logical surface to some primitive revelation or deposit of faith. The appeal may be to any one of a number of primary authoritative positions: to the memory of a founder (as in Zoroastrianism), or a prophet (Moses in Judaism), or to ancient Scriptures (*e.g.,* the Veda and *Upaniṣads* in Hinduism), or an exemplary event (as in Gautama, the Buddha's "enlightenment"), or to God's self-disclosure (as in the Torah, or Law, for Judaism, or in Jesus Christ in Christianity, or Muḥammad's revelations to Islām). Here again, the diversity between doctrines ("allowable interpretations") and the stability of dogmas ("essential teaching") points to the vexed problem of doctrinal development in history that is apparent in all the traditions.

DEVELOPMENT

Every religion has a history of doctrine that is more than a replication of the deposit of faith. Doctrine, as a mode of pedagogy, is conservative of its tradition; as a mode of inquiry, it may be innovative, generating new insights that alter the rhetoric of conventional teaching and, sometimes, its substance as well. There are, of course, wide variations. The persistent continuities between ancient Zoroastrianism and its modern form, Parsiism, or in Jainism, are clearer than those between primitive Hinduism and modern Vedānta (a Hindu philosophical system). All forms and sects of Buddhism appeal jointly to the Three Jewels (the Buddha; the *dharma,* or law; and the *saṅgha,* or monastic order) but are irreconcilable in their differences of interpretation and practice. In each case, the question as to what constitutes legitimate development (*e.g.,* the rival claims of Theravāda, or "Way of the Elders," and Mahāyāna, or "Greater Vehicle," in Buddhism) is left undetermined.

All Jews profess devotion to Torah, even in their disagreements over its authentic observance. Christians profess a common loyalty to the Bible and a common acceptance of the twin dogmas of the Trinity (that the one God is three Persons—Father, Son, and Holy Spirit) and the God-Manhood of Jesus (that Christ is both divine and human) but then divide in their doctrinal systems as they have developed historically. Later dogmas (*e.g.,* transubstantiation, the teaching that the substance of the bread and wine in the Lord's Supper is changed into the substance of the body and blood of Christ, with the properties of the bread and wine remaining unchanged) were defined by the Latin Church without concurrence from Eastern Orthodoxy; the modern dogmas of the Roman Catholic Church (*i.e.,* the immaculate conception of the Virgin Mary, the bodily assumption into heaven of the Virgin Mary, and papal infallibility) were defined in separation from both the Eastern and Protestant consensus. Protestantism has continued an emphasis on its distinctive dogmas of "grace alone" (*sola gratia*), "faith alone" (*sola fide*), and "Scripture alone" (*sola Scriptura*) but has nevertheless undergone immense change and proliferation.

Islām lays great stress on doctrinal stability that is focussed in the Qur'ān, the *sunnah* (custom or tradition), and the consensus (*ijmāʿ*) of its jurists (*'ulamāʾ*). Even so, it has produced doctrinal variants—especially in the medieval period—that are as disparate as the mysticism of the Iranian-born philosopher al-Ghazālī and the rationalism of the Spanish philosopher Averroës and the Persian philosopher Avicenna.

Views of doctrinal development

The process of doctrinal development has been explained variously as a process of logical unfolding or of organic growth, or else as a process of purgations of error and restorations of the original deposit. The notion of a logical unfolding assumes that all that has developed in a religious tradition over the course of its history was already implicit in its original foundation and subsequently had only to become more fully understood. In the case of the doctrine of the Trinity in Christianity, for example, it is argued that the abundant references in the New Testament and the earliest liturgies to God as Father, Son, and Holy Spirit required the development of a dogma that would make explicit the essential Christian trinitarian conviction. Similarly, the dogma on the nature of Christ is understood as the logical outcome of sustained reflection on the testimony about Jesus as the Christ in the Bible and in the apostolic tradition. In the notion of logical unfolding, even in its continual development, truth remains forever unchanged.

Theories of organic development stress the fact that the history of doctrine includes more than explicit formulation of implicit revelation. Such theories take into account the ways in which religious thought is affected by "contemporary" science, philosophy, and historical crises (*e.g.,* the "Copernican revolution" in astronomy, the Renaissance, and other such events). The holders of this view are convinced, however, that all such historical supplementations have been integrated into the original deposit and thus exhibit the power of the religious organization (*e.g.,* the church) to grow and change without substantial alteration of its identity. Thus the 19th-century Roman Catholic cardinal J.H. Newman, in his *Essay on the Development of Christian Doctrine* (1845), argued that

> . . . the highest and most wonderful truths, though communicated to the world once for all by inspired teachers could not be comprehended all at once by the recipients, but, . . . have required only the longer time and deeper thought for their full elucidation (*Introduction,* pp. 29–30).

Newman also believed that this process was safeguarded by the authority of the teaching that would even allow for revisions and occasional corrections of antecedent.

Protestants, by and large, have been more impressed by the lapses and deviations they see in church history and doctrine and thus have tended to construe authentic "development" in terms of a perennial recourse to Scripture and apostolic tradition. Such a view takes historical flux for granted and is less sensitive to the problem of historical continuity.

In all traditions, the course of doctrinal development is crucially affected by the occasional emergence of profound and powerful thinkers who have gathered up scattered elements in their various traditions in freshly relevant syntheses, altering thereby the subsequent history of that tradition. This can be seen, for example, in the North African theologian Augustine's contributions to the making of Latin Christianity and in the matching services of St. John of Damascus in Eastern Orthodoxy. Such also was

the role and contribution of Moses Maimonides in medieval Judaism (*e.g.,* the Thirteen Articles of Faith in his commentary on the Mishna) and of St. Thomas Aquinas in medieval Christianity (*e.g., Summa theologiae*). The 16th-century Reformers Martin Luther and John Calvin gave Protestantism its classical form, to be followed by yet other and different system builders (*e.g.,* Friedrich Schleiermacher in the 19th century and Karl Barth in the 20th century).

Each theory of development has had its own distinctive prescription for doctrinal stability and doctrinal change. In Christianity, Eastern Orthodoxy locates its authority in "Holy Tradition," which is fixed and guided by the dogmas proclaimed by the ecumenical councils. Roman Catholicism relies on the magisterium (teaching authority) of the church, which is directed by the bishops as a "college" (*collegium episcoporum*) and supremely by the bishop of Rome as their collegial head. Protestantism has sought to bind both tradition and the church to the authority of Holy Scripture, with the resulting problem of specifying what is to be regarded as truly authoritative interpretations of Scripture.

THE RELATION OF FAITH, REASON, AND RELIGIOUS INSIGHT TO DOCTRINE AND DOGMA

The interplay of faith and reason in religious experience

Insofar as doctrines and dogmas represent conceptualizations of the human encounter with the divine mystery, they are bound to reflect the interplay of faith and reason in religious experience and to imply some notion of levels and stages in the progress of believers as they move from the threshold of faith toward its fulfillment. Doctrine is concerned with communication and consensus, with the exposure of the religious vision to rational probes and queries. There is, therefore, a tension in all religions between mystical intuition and logical articulation, between insight and dialogue. Most traditions agree that perfect understanding is a goal that lies beyond a "simple faith" and the routine observance of rites and duties. Most of them also agree that the utmost pinnacle of religious insight is ineffable. One mode of differentiation between doctrinal traditions, therefore, is their relative openness or resistance to the auxiliary services of philosophy and science of faith's fulfillment.

In the majority of the religions of the East, very broadly, reason's chief role is the purgation of illusion and self-deception so that souls may follow the ways of wisdom and right conduct to their true fulfillment in Nirvāṇa. The Hindu passes from the initiatory level of "the student" (who is dependent on a teacher, guru) to the ambivalent freedom of "the householder," to the great freedom of "the forest dweller," to the fullest freedom of the *sannyāsin* (an enlightened ascetic). Reason, which is chiefly reflective, assists at every stage in perfecting faith's self-understanding. In Buddhism, one follows the *dharma* from *saddha* (which is practical knowledge of one's religious obligations), to *jñāna* (rational insight), to *vijñāna* (mystical illumination).

In most of the religions of the West, again very broadly, the primary function of reason is seen to be that of rendering the mysteries of faith as intelligible as possible, in support of the intellectual love of God. In Judaism, progress in the knowledge of Torah is focussed in the Bible and the Talmud (commentaries on the Law), guided by the twin hermeneutical (critical interpretive) principles of Halakha (the oral precepts and decisions of the rabbis) and of Haggada (instructive stories, parables, and other similar devices).

Variations in Islām range from the rigid orthodoxy of the Ḥanbalites (a conservative school of law following the teachings of Ibn Ḥanbal), to the rational liberalism of the Mu'tazilites (a school of law utilizing Greek philosophical methods), to the dialectical doctrines (*kalām*) of the Arabian theologian al-Ash'arī and the Turkish philosopher al-Fārābī. All of these, however, are anchored in the twin dogmas of the unity of God and the prophetic office of Muhammad. Spiritual progress is measured by the believer's faithfulness in obedience to the "Five Pillars," or religious duties, including prayer, fasting, and the pilgrimage to Mecca.

In Christianity, the dialectic between faith and reason has ranged from the fideism (emphasis on faith) of the 2nd-century North African theologian Tertullian to the intellectualism of Thomas Aquinas. An ancient distinction between faith as bare assent to orthodox doctrine (*fides informis*) and faith as existential trust in God's grace (*fides formata*) gave rise to the further distinction between faith as a set of doctrines to be believed (*fides quae creditur*) and faith as personal involvement (*fides qua creditur*). Philipp Melanchthon, a 16th-century Lutheran Reformer, stressed the point that even the devils are "orthodox" (having "dead faith") but to no avail, since only those who have embraced God's reconciling love (*fiducia*) receive the benefits of salvation ("living faith"). In general, this distinction has become standard in Protestantism.

Faith as assent and as trust

CHANGING CONCEPTIONS

In all the great religious traditions, and between them, the clash of doctrines and dogmas has, more often than not, been polemical. The *odium theologorum* ("bitterness of the theologians") of which Melanchthon once complained so plaintively has been notorious. Within the several traditions, doctrinal disputes have sometimes led to division or else have accompanied divisions caused otherwise. In relationships between the great world religions, dogmas and doctrines have usually been regarded as mutually exclusive. There are, however, significant signs of change in this attitude. The rise and spread of the ecumenical movement in the 20th century and notable advances in the comparative study of world religions reflect an enlarged commitment to the widest possible community of mutual religious interests. The "Decree on Ecumenism" and its "Declaration on the Relationship of the Church to Non-Christian Religions" of the Roman Catholic second Vatican Council (1962–65) are signal instances of this new disposition.

(A.C.O.)

MAJOR THEMES AND MOTIFS

Creation

Doctrines of creation are philosophical and theological elaborations of the primal myth of creation within a religious community. The term myth here refers to the imaginative expression in narrative form of what is experienced or apprehended as basic reality (see also MYTH AND MYTHOLOGY). The term creation refers to the beginning of things, whether by the will and act of a transcendent being, by emanation from some ultimate source, or in any other way.

NATURE AND SIGNIFICANCE

The myth of creation is the symbolic narrative of the beginning of the world as understood by a particular community. The later doctrines of creation are interpretations of this myth in light of the subsequent history and needs of the community. Thus, for example, all theology and speculation concerning creation in the Christian community are based on the myth of creation in the biblical book of Genesis and of the new creation in Jesus Christ. Doctrines of creation are based on the myth of creation, which expresses and embodies all of the fertile possibilities for thinking about this subject within a particular religious community.

The basis for man's placement and orientation in the world

Myths are narratives that express the basic valuations of a religious community. Myths of creation refer to the process through which the world is centred and given a definite form within the whole of reality. They also serve as a basis for the orientation of man in the world. This centring and orientation specify man's place in the universe and the regard he must have for other humans, nature, and the entire nonhuman world; they set the stylistic tone that tends to determine all other gestures, actions,

and structures in the culture. The cosmogonic (origin of the world) myth is the myth par excellence. In this sense, the myth is akin to philosophy, but, unlike philosophy, it is constituted by a system of symbols; and because it is the basis for any subsequent cultural thought, it contains rational and nonrational forms. There is an order and structure to the myth, but this order and structure is not to be confused with rational, philosophical order and structure. The myth possesses its own distinctive kind of order.

Myths of creation have another distinctive character in that they provide both the model for nonmythic expression in the culture and the model for other cultural myths. In this sense, one must distinguish between cosmogonic myths and myths of the origin of cultural techniques and artifacts. Insofar as the cosmogonic myth tells the story of the creation of the world, other myths that narrate the story of a specific technique or the discovery of a particular area of cultural life take their models from the stylistic structure of the cosmogonic myth. These latter myths may be etiological (*i.e.*, explaining origins); but the cosmogonic myth is never simply etiological, for it deals with the ultimate origin of all things.

The ritual expression of creation: sacred time

The cosmogonic myth thus has a pervasive structure; its expression in the form of philosophical and theological thought is only one dimension of its function as a model for cultural life. Though the cosmogonic myth does not necessarily lead to ritual expression, ritual is often the dramatic presentation of the myth. Such dramatization is performed to emphasize the permanence and efficacy of the central themes of the myth, which integrates and undergirds the structure of meaning and value in the culture. The ritual dramatization of the myth is the beginning of liturgy, for the religious community in its central liturgy attempts to re-create the time of the beginning.

From this ritual dramatization the notion of time is established within the religious community. To be sure, in most communities there is the notion of a sacred and a profane time. The prestige of the cosmogonic myth establishes sacred or real time. It is this time that is most efficacious for the life of the community. Dramatization of sacred time enables the community to participate in a time that has a different quality than ordinary time, which tends to be neutral. All significant temporal events are spoken of in the language of the cosmogonic myth, for only by referring them to this primordial model will they have significance.

In like manner, artistic expression in archaic or "primitive" societies, often related to ritual presentation, is modelled on the structure of the cosmogonic myth. The masks, dances, and gestures are, in one way or another, aspects of the structure of the cosmogonic myth. This meaning may also extend to the tools man uses in the making of artistic designs and to the precise technique he employs in his craft.

Mention has been made above of the fact that the cosmogonic myth situates mankind in a place, in space. This centring is at once symbolic and empirical: symbolic because through symbols it defines the spatiality of human beings in ontological terms (of being) and empirical because it orients them in a definite landscape. Indeed, the names given to the flora and fauna and to the topography are a part of the orientation of humans in a space. The subsequent development of language within a human community is an extension of the language of the cosmogonic myth.

The primordial structure of culture

The initial ordering of the world through the cosmogonic myth serves as the primordial structure of culture and the articulation of the embryonic forms and styles of cultural life out of which various and differing forms of culture emerge. The recollection and celebration of the myth enable the religious community to think of and participate in the fundamentally real time, space, and mode of orientation that enables them to define their cultural life in a specific manner.

TYPES OF COSMOGONIC MYTHS

The world as a structure of meaning and value has not appeared in the same manner to all human civilizations. There are, therefore, almost as many cosmogonic myths as there are human cultures. Until quite recently, the classification of these myths on an evolutionary scale, from the most archaic cultures to contemporary Western cultures (*i.e.*, from the assumedly simplest to the most complex) was the most dominant mode of ordering these myths. Recent 20th-century scholars, however, have begun to look at the various types of myths in terms of the structures that they reveal rather than considering them on an evolutionary scale that extends from the so-called simple to the complex, for, in a sense, there are no simple myths regarding the beginning of the world. The beginning of the world is simultaneously the beginning of the human condition, and it is impossible to speak of this beginning as if it were simple.

Creation by a supreme being. The 19th-century scholars who took an evolutionary survey of human culture and religion (*e.g.*, Sir James George Frazer and Edward Burnett Tylor) held that the notion of the creation of the world by a supreme being occurred only in the highest stage of cultural development.

Primordial supreme beings or creator gods

Andrew Lang, a Scottish folklorist, challenged this conception of the development of religious ideas, for he found in the writings of anthropologists, ethnologists, and travellers evidence of a belief in a supreme being or high god among cultures that had been classified as the most primitive. This position was taken up and elaborated by an Austrian priest-anthropologist, Wilhelm Matthäus Schmidt, who reversed the evolutionary theory, holding that there was a primordial notion of a supreme being, a kind of original intellectual and religious conception of a single creator god, that degenerated in subsequent cultural stages. Though Schmidt's theories of cultural historical stages and diffusion and an original primordial revelation have for the most part been discredited and abandoned, the existence of a belief in a supreme being among primitive peoples (a notion discovered by Andrew Lang) has been proven and attested to over and over again by investigators of numerous cultures. This belief has been found among the cultures of Africa, the Ainu of the northern Japanese islands, Amerindians, south central Australians, the Fuegians of South America, and in almost all parts of the globe.

Though the precise nature and characteristics of the supreme creator deity may differ from culture to culture, a specific and pervasive structure of this type of deity can be discerned. The following characteristics tend to be common: (1) he is all wise and all powerful. The world comes into being because of his wisdom, and he is able to actualize the world because of his power. (2) The deity exists alone prior to the creation of the world. There is no being or thing prior to his existence. No explanation can therefore be given of his existence, before which one confronts the ultimate mystery. (3) The mode of creation is conscious, deliberate, and orderly. This again is an aspect of the creator's wisdom and power. The creation comes about because the deity seems to have a definite plan in mind and does not create on a trial-and-error basis. In Genesis, for example, particular parts of the world are created seriatim; in an Egyptian myth, Kheper, the creator deity, says, "I planned in my heart," and in a Maori myth the creator deity proceeds from inactivity to increasing stages of activity. (4) The creation of the world is simultaneously an expression of the freedom and purpose of the deity. His mode of creation defines the pattern and purpose of all aspects of the creation, though the deity is not bound by his creation. His relationship to the created order after the creation is again an aspect of his freedom. (5) In several creation myths of this type, the creator deity removes himself from the world after it has been created. After the creation the deity goes away and only appears again when a catastrophe threatens the created order. (6) The supreme creator deity is often a sky god, and the deity in this form is an instance of the religious valuation of the symbolism of the sky.

In creation myths of the above type, the creation itself or the intent of the creator deity is to create a perfect world, paradise. Before the end of the creative act or sometime soon after the end of creation, the created order or the intent of the creator deity is thwarted by some fault of one

Rupture of primordial harmony and perfection

of the creatures. There is thus a rupture in the creation myth. In some myths this rupture is the cause of the departure of the deity from creation.

An African myth from the Dogon peoples of West Africa illustrates this point. In this myth the creator deity first creates an egg. Within the egg are two pairs of twins, each pair consisting of one male and one female. These twins are supposed to mature within the egg, becoming at maturation androgynous (both male and female) beings, the perfect creatures to inhabit the earth. One of the twins breaks from the egg before maturation because he wishes to dominate the creation. In so doing he carries a part of the egg with him, and from this he creates an imperfect world. The creator deity, seeing what he has done, sacrifices the other twin to establish a balance in the world. The creation is sustained by this sacrifice, and it is now ambiguous, instead of the perfect world intended by the god.

This myth not only shows how a rupture takes place within the myth itself but also points out the fact that the characteristics of the supreme creator deity noted above seldom exist apart from other mythological contexts. The widespread symbols of dualism (the divine twins), the cosmic egg, and sacrifice are basic themes in the structure of this African creation myth. In myths of this kind, however, prominence must always be given to the might of a powerful creator sky deity under whose aegis the created order comes into being.

Creation through emergence. In contrast to the creation by a supreme sky deity, there is another type of creation myth in which the creation seems to emerge through its own inner power from under the earth. In this genre of myth, the created order emerges gradually in continuous stages. It is similar to a birth or metamorphosis of the world from its embryonic state to maturity. The symbolism of the earth or a part of the earth as a repository of all potential form is prominent in this type of myth. In some myths of this type (*e.g.*, the Navajo myth of emergence), the movement from a lower stage to a higher one is initiated by some fault of the people who live under the earth, but these faults are only the parallels of an automatic upper movement in the earth itself.

From the embryonic or inchoate to the mature and definite

Just as the supreme-creator-deity myth forms a homology to the sky, the emergence myth forms a homology to the earth and to the childbearing woman. In many cases the emergence of the created order is analogous to the growth of a child in the womb and its emission at birth. This symbolism is made clear in a Zuni myth that states,

> Anon is the nethermost world, the seed of men and creatures took form and increased; even as in eggs in warm places speedily appear . . . Everywhere were unfinished creatures, crawling like reptiles one over another, one spitting on another or doing other indecencies . . . until many among them escaped, growing wiser and more manlike.

The underworlds prior to the created order appear chaotic; the beings inhabiting these places seem without form or stability, or they commit immoral acts. The seeming chaos is moving toward a definite form of order, however, an order latent in the very forms themselves rather than from an imposition of order from the outside.

From another perspective the emergence myth is homologous to the seed. When the homologue of the seed is referred to, the meaning of fertility and death are at once introduced. The seed must die before it can be reborn and actualize its potentiality. This symbolism is dramatically presented in a wide range of funerary rites: one is buried in the earth in hope of a renewal from the earth, or the earth is the repository of the ancestors from whom the new generation emerges. In every case, emergence myths demonstrate the latent potency immanent in the earth as a repository of all life forms.

Creation by world parents. Closely related to the above type of myth is the myth that states that the world is created as the progeny of a primordial mother and father. The mother and father are symbols of earth and sky, respectively. In myths of this kind, the world parents generally appear at a late stage of the creation process; chaos in some way exists before the coming into being of the world parents. In the Babylonian myth *Enuma elish*, it is stated,

> When on high the heaven had not been named
> Firm ground below had not been called by name,
> Naught but primordial Apsu, their begetter,
> (And) Mummu-Tiamat, she who bore them all,
> Their waters comingling as a single body;

The Maori make the same point when they state that the world parents emerge out of *po*. *Po* for the Maori means the basic matter and the method by which creation comes about. There is thus some form of reality before the appearance of the world parents.

From inertness to activity

Even though the world parents are depicted and described as in sexual embrace, no activity is taking place. They appear as quiescent and inert. The chthonic (underworld) structure of the earth as latent potentiality tends to dominate the union. The parents are often unaware that they have offspring, and thus a kind of indifference regarding the union is expressed. The union of male and female in sexual embrace is another symbol of completeness and totality. As in the African myth from the Dogon referred to above, sexual union is a sign of androgyny (being both male and female) and androgyny, in turn, a sign of perfection. The indifference of the world parents is thus not simply a sign of ignorance but equally of the silence of perfection. The world parents in the Babylonian and Maori myths do not wish to be disturbed by their offspring. As over against the parents, the offspring are signs of actuality, fragmentation, specificity; they define concrete realities.

The separation of the world parents is again a rupture within the myth. This separation is caused by offspring who wish either to have more space or to have light, for they are situated between the bodies of the parents. In some myths the separation is caused by a woman who lifts her pestle so high in grinding grain that it strikes the sky, causing the sky to recede into the background, thus providing room for the activities of mankind. In both cases an antagonistic motive must be attributed to the agents of separation. In the Babylonian and Maori versions of this myth, actual warfare takes place as a result of the separation.

Over against the primordial union of the world parents, there is the desire for knowledge and a different orientation in space. After the separation, lesser deities related to solar symbolism take precedence in the creation. The sun and light must be seen in these myths as representing the desire for a humanizing and cultural knowledge as over against the passive and inert forms of the union of the parent deities. From the point of separation, the mythic narrative of the world-parent myths states how different forms of cultural knowledge are brought to man by the offspring, the agents of separation. The separation of the world parents is the sign of a new cosmic order, an order dedicated to the techniques, crafts, and knowledge of culture.

Creation from the cosmic egg. In the Dogon myth referred to above, the creation deity begins the act of creation by placing two embryonic sets of twins in an egg. In each set of twins is a male and female; during the maturation process they are together thus forming androgynous beings. In a Tahitian myth, the creator deity himself lives alone in a shell. After breaking out of the shell, he creates his counterpart, and together they undertake the work of creation.

A Japanese creation narrative likens the primordial chaos to an egg containing the germs of creation. In the Hindu tradition the creation of the world is symbolized in the Chandōgya *Upaniṣad* by the breaking of an egg, and the universe is referred to as an egg in other sources. The Buddhists speak of the transcending of ordinary existence, the realization of a new mode of being, as breaking the shell of the egg. Similar references to creation through the symbol of the egg are found in the Orphic texts of the Greeks and in Chinese myths.

The egg is a symbol of the totality from which all creation comes. It is like a womb containing the seeds of creation. Within the egg are the possibilities of a perfect creation (*i.e.*, the creation of androgynous beings). The egg, in addition to being the beginning of life, is equally a symbol of procreation, rebirth, and new life. In a version

of the Dogon, one of the twins returns to the egg in order to resuscitate the other.

The cosmogonic water and the diver animal or diver devil

Creation by earth divers. Two elements are important in myths of this type. There is, first, the theme of the cosmogonic water representing the undifferentiated waters that are present before the earth has been created. Secondly, there is an animal who plunges into the water to secure a portion of earth. The importance of the animal is that the creature agent is a prehuman species. This version of the myth is probably the oldest version of this genre. This basic structure of the earth-diver myth has been modified in central Europe in myths that relate the story of the primordial waters, God, and the devil. In these versions of the earth-diver myth, the devil appears as God's companion in the creation of the world. The devil becomes the diver sent by God to bring earth from the bottom of the waters. In most versions of this myth, God does not appear to be omniscient or omnipotent, often depending on the knowledge of the devil for certain details regarding the creative act—details that he learns through tricks he plays upon the devil.

In still different versions of this myth, the relationship between God and the devil moves from companionship to antagonism; they become adversaries, though they remain as co-creators of the world. The fact that the devil has had a part in the creation of the world is one way of explaining the origin and persistence of evil in the world.

Mircea Eliade, a noted 20th-century historian of religions, has pointed to another theme in certain Romanian versions of this myth. After God has instructed the devil to dive to the bottom of the waters and bring up the earth, the devil obeys, diving several times before he is able to bring up and hold on to a small portion of earth. After the creation of the world from this small portion of earth, God sinks into a profound sleep. This sleep is a sign of mental exhaustion, for only the devil and a bee know the solution to certain details of the creation, and God must, with the help of the bee, trick the devil into giving him this vital information. God's sleep, according to Eliade, is a sign of his passivity and disinterest in the world after it has been created, and it harks back to certain archaic myths in which the supreme deity retires from the world after its creation, becoming disinterested and passive in the relationship to his work.

God's post-creation sleep

DOCTRINES OF CREATION

Some of the major types of creation myths have been presented above. It is from myths of this sort and their dominant themes that theological and philosophical speculation have been developed in the various religious communities throughout the world.

Basic mythical themes. *Primordiality.* In several myths it is stated that the primordial stuff of creation was some form of undifferentiated matter (*e.g.*, water, chaos, a monster, or an egg). It is from this undifferentiated matter that the world evolves or is made. In the case of the egg and monster symbols, there seems to be a notion of a definite original form, but the egg is undifferentiated; for its form is vague and embryonic, and the monster figure—containing all of the forms of chaos in a terrible way—expresses the theme that chaos is not only passive (as is water) but resists creation. Although creation results as a modification of the primordial matter, however, it is this matter that determines and sets the limits to the extension of the world in space and time. Thus, in communities in which myths of this type find their expression, there are periods of mythical-ritual renewal at certain cyclical periods in which the world returns to its original chaos to rise again out of this initial state.

The creator deity as sole determiner of creation's nature and destiny

When it is stated that the supreme being created the world and that there was no primordial matter prior to his being, then the determination of the world is in the mind and will of the deity. This leads to distinctive conclusions regarding the destiny of the world and man. The end (and meaning) of the world is thus not determined by the primordial matter but by the deity who created the world. It is he alone who determines the preservation, maintenance, and end of the world.

Dualisms and antagonisms. In emergence myths there seems to be an easy movement from one stage of creation to the next, but, as has been shown in the Navajo myth, at each subterranean level there is some type of antagonism among the developing embryonic creatures. This is one of the reasons for the separation of the creatures and the movement to another level. Though the emergence myths portray the mildest form of this antagonism, it is still present in myths of this sort.

In the world-parent myths there is antagonism between the offspring and the parents. This is a conflict between generations, expressing the desire of the children to determine their own place and orientation in existence against the passivity of the parents.

A dualism and antagonism is found again in the cosmic-egg myths, especially in the myths in which the egg contains twins. One twin wishes to take credit for the creation of the world alone, interrupting the harmonious growth within the egg before maturation. The faulty creation by this evil twin accounts for the ambiguous nature of the world and the origin of evil.

This observation applies equally to the dualistic structure in some versions of the earth-diver myths. The devil moves in the various versions of this myth from the companion to the antagonist of God, possessing the power to challenge the deity.

Sacrifice of primordial beings

Creation and sacrifice. In many cosmogonic myths, the narrative relates the story of the sacrifice and dismemberment of a primordial being. The world is then established from the body of this being. In the myth *Enuma elish,* the god Marduk, after defeating Tiamat, the primeval mother, divides the body into two parts, one part forming the heavens, the other, the earth. In a West African myth, one of the twins from the cosmic egg must be sacrificed to bring about a habitable world. In the Norse *Prose Edda,* the cosmos is formed from the body of the dismembered great Ymir, and, in the Indian Rgveda, the cosmos is a result of the sacrifice of man.

In these motifs of sacrifice, something similar to the qualification of the undifferentiated matter of creation is suggested, for, just as the primal stuff of creation must be differentiated before the world appears, the sacrifice of primordial beings is a destruction of the primal totality for the sake of a specific creation.

When the victim of the sacrifice is a primal monster, the emphasis is on the stabilization of the creation through the death of the monster. The monster symbolizes the strangeness and awesomeness occurring when a new land or space is occupied. The "monster" of the place is the undifferentiated character of the space and must be immobilized before the new space can be established.

In a myth from Ceram (Molucca Islands), a beautiful girl, Hainuwele, has grown up out of a coconut plant. After providing the community with their necessities and luxuries, she is killed and her body cut into several pieces, which are then thrown over the island. From each part of her body a coconut tree grows. It is only after the death of Hainuwele that mankind becomes sexual; that is, the murder of Hainuwele enables mankind to have some determination in the process of bringing new life into the world.

Theological and philosophical doctrines. Myths and poetic renderings in legends, sagas, and poetry express the basic cultural insights into some of the elements involved in the human consciousness about creation. Theological, philosophical, and scientific theory are types of rationalizations of these basic insights in terms of the particular culture and historical periods of the cultures in question.

The attempt to integrate the meanings of primordiality, dualisms and antagonisms, sacrifices, and ruptures and to meet demands of some kind of logical order and, at the same time, keep alive the meaning of these structures as religious realities, objects of worship, and a charter for the moral life, has led to the development of doctrines.

In "primitive" and archaic societies, the correct ritual enactment of mythical symbols ensures the order of the world. These rituals usually take place at propitious moments (*e.g.*, at the birth of a child, marriage, the founding of a new habitation, the erection of a house or temple, the beginning of a new year). In each case, the seemingly

practical activities imitate the mythic structure of the first beginning.

Theological and philosophical speculations and controversies centre within and between religious communities over the issues of the primordial nature of reality, dualisms, the process of creation, and the nature of time and space. A doctrine of creation must contain or suggest the manner in which all cultural meanings, both empirical and abstract, constitute an integral totality. Speculations that are based on the initial insights of a mythical theme explicate some principle in the myth as a basis for generalization and logical form on which all elements and themes may be ordered.

Transcendence and otherness. Doctrinal positions may be modelled around any or all of the themes of the cosmogonic myth. If the emphasis falls upon creation by a high god through his thought, word, or other mode, the problem of the otherness and difference between creator and creature becomes a source of theological discussion and philosophical speculations. In Judaism, Christianity, and Islām, the classical locus of this issue is found. All of these religions have theological traditions that raise this problem. Related to this issue is the transcendence and arbitrary action of the creator deity. Because he is prior to the world and its creatures, the question arises whether there are modes of creaturely knowledge or apprehension that are capable of knowing him; of whether he is subjected to the same categories of being as his creatures; of whether his time and space are the same time and space of his creation.

The problem of difference between creator and creature

To some extent, the a priori nature of this type of deity creates an apparent dualism between the creator and the world and creatures. This dualism is mediated in various forms in the traditions. In Judaism it is mediated through nature and the covenant Yahweh has with his people; in Christianity through the mediatorship of his son, Jesus Christ; and in Islām through the sacred word of the Qur'ān by the prophet Muḥammad. Even within these traditions, however, the transcendent nature of the deity and his mediatorship through some other being or principle does not settle the doctrinal issue, for different cultural-historical periods of these traditions offer a variety of theological speculation concerning the nature and meaning of the deity, the world, and the mediator. The traditions offer a structure through which such speculation is ordered and clarified.

Creation through emanations. The theme of emergence is related to theological and philosophical notions of emanations from a single principle and the idea of the transmutation of being. Ideas of this kind are found in "primitive" religion (Dogon, Polynesian), in Taoism, and in the Pre-Socratic philosophers Thales and Anaximander.

In one version of the Dogon myth, creation proceeds from a small seed. Within the seed spontaneous movements begin. These movements, which burst from the shell of the seed and make contributions in space, create all forms of beings and the universe. Similarly, in the Polynesian myth Ta-aroa develops the world out of himself and the shell in which he lived.

A pervasive theme in Chinese thought is that of a universe in a perpetual flux. This flux follows a fixed and predictable pattern either of eternal oscillation between two apparently opposed poles or of a cyclical movement in a close orbit. The oscillation pattern is expressed by the Yin-Yang doctrine of Taoism. In the five element doctrine, a cyclical movement is correlated with the five elements, earth, wood, metal, fire, and water: these in turn form an equivalence with the third month of summer and with spring, autumn, summer, and winter, respectively. These parallelisms then form equivalences with the five directions, and they in turn with the five primary colors. Ancient Chinese thinkers never discuss an initial conscious act of creation. The cyclical movement itself produced the empirical and abstract form of the cosmos. The oscillation between the Yin and the Yang forms a correlation in all phenomena extending to the realms of time, space, number, and ethics.

The notion of the universe in a perpetual flux

Thales thought that the fundamental principle of cosmos was water. The earth floated on water; water was the natural cause of all things. Anaximander taught that there was an eternal undestructible something out of which everything arises and everything returns. In other words, the fundamental substratum of the world could not be an element of the world. The importance of Anaximander was in his use of the term *archē* ("beginning" or "rule") to refer to a principle unlike any other principle or element in the world to explain the cause of all other things in the universe.

Dualisms. Dualistic conceptions of creation come to the fore in the theme of earth-diver myths, in which there is an antagonism between the co-creators of the universe. This conception is present again in myths of divine twins and in Zoroastrianism where the Ormazd and Ahriman represent the creative and destructive principles in creation. In some sense this is not an ontological dualism for the first creative act of Ormazd was the limitation of time and thus the limitation of the power of Ahriman to carry out his destruction. Doctrines of this kind are related to the origin of evil in the world.

SKEPTICISM REGARDING CREATION

Alongside the various myths and doctrines regarding creation, there are equally skeptic positions concerning the unknowability of creation. This critique is present in several religious and philosophical traditions. It may be correlated with the mythical meaning of *deus otiosus,* the deity who retires from the world after his creation, or with the mythic theme from some earth-diver myths that emphasize the physical and intellectual fatigue of the deity after creation. In the first case, the removal of the deity from creation leaves no access to his plan or will; in the other case, because of the fatigue of the deity who has exhausted all of his knowledge in creation, there is thus nothing for man to learn from him.

In the Indian tradition the Rigveda, an ancient sacred text, expresses skepticism in this manner:

> He, the first origin of this creation, whether he formed it all or did not form it,
> Whose eye controls this world in highest heaven, he verily knows it, or perhaps he knows not.

The Buddha declared certain cosmological and metaphysical questions unanswerable. His refusal to answer questions of this kind gave rise to the "silence of the Buddha" as a philosophical style in Buddhism. They included such questions as: whether the world is eternal or not or both; whether the world is finite (in space) or infinite or both or neither.

In the Chinese tradition Kuo Hsiang (died AD 312) questioned the origin of the basic oscillation of the Taoist movement. For Hsiang there is no such thing as Non-Being for Being is the only reality. Being could not have evolved from Non-Being nor can it revert to Non-Being. As Kuo Hsiang put it,

> I venture to ask whether the Creator is or is not? If He is not, how can He create things? If He is, then (being one of these things), He is incapable of creating the mass of bodily forms. . . . The creating of things has no Lord; everything creates itself. Everything produces itself and does not depend on anything else. This is the normal way of the universe.

Skepticism of this same kind is expressed by Parmenides, a Pre-Socratic, and in the modern tradition of Western philosophy from Immanuel Kant's *Kritik der reinen Vernunft* (1st ed. 1781; Eng. trans., *Critique of Pure Reason,* 1929) to Ludwig Wittgenstein's *Tractatus Logico-Philosophicus* (1922). Skepticism of this kind about the nature of the cosmic order and especially about the ultimate origin of the universe places limitations on the possibility of the rational consciousness to authentically ask these questions. In some instances theologians have agreed and held to a notion of revelation as a response to these unanswerable questions. In other cases, the questions themselves have been labelled nonsensical.

The work of Hartshorne and Reese

Charles Hartshorne and William Reese, 20th-century U.S. philosophers, have attempted to clarify and criticize all possible rational reflections concerning the relationship of deity to the universe. They state two opposed positions. The first is that of classical theism in which there is the admission of plurality, potentiality, becoming, as a

secondary form of existence outside of God. The other position, that of classical pantheism, says that though God includes all within himself, he cannot be complex or mutable, for such categories only express human ignorance and illusion. They attempt to overcome this dilemma by combining these contrary poles into a dipolar conception of the meaning of deity. Because classical theism is primarily a Western approach to the problem and classical pantheism an Eastern approach, the dipolar conception is at the same time a synthesis of Western and Eastern thought. In addition to this, these philosophers set forth a method of analyzing all conceptions of deity and world according to basic religious and rational categories. As metaphysicians they go far in refuting the skepticism regarding rational knowledge of the relationship between the deity and the universe. (C.H.Lo.)

Eschatology

Eschatology (the doctrine of last things) is originally a Western term, referring to Jewish and Christian beliefs about the end of history (or of the world in its present state), the resurrection of the dead, the Last Judgment, and related matters. The term has been extended by historians of religions to cover similar themes and concepts in the religions of nonliterate peoples, ancient Mediterranean and Middle Eastern cultures, and Eastern civilizations.

Eschatological ideas and beliefs played a central role in the development of Judaism; the Kingship of God, "the end of days," "the world to come," the Messiah and the messianic era, the Day of Judgment, and the images of a perfected future were basic concerns in biblical (Old Testament) and rabbinic Judaism.

NATURE AND SIGNIFICANCE

In New Testament Christianity, history is viewed throughout in eschatological terms: the future of God has already begun with the appearance of Christ; the end of history is near; the end time is therefore filled with danger and salvation, faith and unfaith, Christ and Antichrist, and will be consummated through the resurrection of the dead, the judgment of the world, and its salvation through a new creation. Christianity, with this biblical heritage, has influenced many religions, revolutions, and civilizations through its orientation toward hope in the future. Biblical eschatological archetypes can be found in the various secular liberation movements leading up to the present day.

In the general history of religions, the term eschatology refers to *conceptions of the beyond* that express the destiny of man after his death (immortality of the soul, rebirth, resurrection, migration of the soul). These eschatological concepts stand in a mutual relationship with the experiences of men in the present world, the turning points of life, and the understanding of death. They often pose a contrast to the present experience of suffering within nature and society, within this whole "perverted world." Eschatological themes thrive particularly in crisis situations, whether they serve as consolations for those who hope for a better world or as the motives for revolutionary transformation of the world.

A distinction has often been made between individual (personal) eschatology and collective (social) eschatology. The eschatological expectations of either type, however, are as extensive as the respective believer's interest and involvement in life and his suffering from its miseries. These expectations can embrace individual souls; they can just as easily embrace a people or group, humanity, or the whole cosmos.

GENERAL CHARACTERISTICS

The theme of origins and last things. Because biblical eschatology is grounded in what are interpreted as uniquely occurring historical events (such as the Exodus of the Hebrews from Egypt in the 13th century BC), certain difficulties occur when the biblical concept of eschatology is translated into the eschatological framework of other religions. In other religions (especially Eastern religions and the religions of nonliterate peoples), cosmic representations of the eternal struggle between cosmos and chaos prevail. Therefore, a distinction must clearly be made between *mythical* eschatologies and *historical* eschatologies. The term mythical refers to the concept of the philosophically conceived truth of the human condition in relation to the realm of the sacred and the profane as defined in nontemporal terms and stories. The term historical, on the other hand, refers to the concept of the philosophically conceived truth of the human condition in relation to the realm of the sacred and the profane as defined in temporal terms and stories.

Mythical eschatologies

Mythical eschatologies emphasize the reproduction of the origin of the world at the end of the world. The end time repeats the primordial time; out of chaos (disorder) there arises anew the cosmos (order). At the origin there stands the cosmogony (creation of order) and the laws that govern the world and the pure order of things. Time is experienced as decay, degeneration, and guilt. Salvation, therefore, is found in a return of the origin or a return to the origin. All historical events are interpreted as representations of an eternal struggle in which the world order is defended against chaos. This struggle occurs as myth in the world of the gods and as history in the world of men. History thus becomes a cultic drama in which priests and kings play out the ritual roles foreordained to them.

The religious symbolism of mythical eschatology can be defined in terms of the "myth of the eternal return." This concept contains not only a cyclical view of history but also a cultic view of the annihilation of the horrors of history itself. In the ever-recurring cultic festivals, the lost time of history is regenerated and eternity is represented. Through the ritualistic repetition of all events in the creation of the cosmos, the impression of transience is proved to be a mere semblance. Everything basically remains in its place. In the framework of the myth of the eternal return, hope is inherent in memory. This basic structure is not limited to the great cultic religions; many messianic and revolutionary movements (such as the nativistic religious movements in Africa and Oceania, and sectarian movements in pre-1917 Russia) also exemplify it. Within a history that is generally regarded as evil, the saving future is depicted as a return to the primordial origin. In the terms of mythical eschatology, the meaning of history is found in a celebration of the eternity of the cosmos and the repeatability of the origin of the world.

Historical eschatology

Historical eschatology, on the contrary, is not grounded in a mythical primal happening but in historically datable events that are perceived as root experiences (past events that are narrative and paradigmatic in the present) and are regarded as fundamental for the progress of history. Biblical and biblically influenced eschatologies are grounded in such a view of historical experiences and are directed toward the historical future inaugurated by them. In this view, such experiences are never universal but always particular; they are not grounded in natural happenings but in historical election. Such events and experiences are not repeatable in cultic forms and rituals but are remembered through the telling of history and the relating of tradition. They do not abolish history; they rather inaugurate a new process (or new age) of history. They are events in which a *novum* (new or extraordinary thing) is perceived, that have a greater future than a beginning. Hope is thus grounded in historical remembrance but goes beyond what is remembered historically.

Because history is viewed as unrepeatable, the future of history is final. History is understood in this context not as the arena of the horrors of chaos but as the field of danger and salvation. The meaning of history is thus not found in its cultic abolishment (as in the case of mythical eschatologies) through the presence of eternity but in its future goal and its fulfillment. The divine or sacred is not experienced in the eternally recurring orders of nature and of the cosmos through ritualistic reenactments. Rather, God's freedom, faithfulness, and promises of the future are known and comprehended in the contingent and irreversible events of history. If here the future is greater than the beginning, hope at the end is more extensive than at the beginning.

Historical eschatologies are found in the faith of Israel and Judaism, which is grounded in the Exodus event (the

liberation of the Hebrews from Egypt in the 13th century BC) and which in the course of its historical experiences is more and more directed towards the expected revelation of the glory of God in all lands. Historical eschatologies are also found in the Christian faith, which is grounded in the history of Jesus and in the root experience of his resurrection from the dead. The hope of Christian faith is aimed at the Kingdom of Christ and the Kingdom of God, through which history is ended as well as fulfilled. In both cases the unique occurrence of a historical foundation event serves as a basis for the final goal of the long-awaited and hoped for future. A historically experienced *novum* opens up hope in a new creation that will be more than the reproduction of the primordial condition.

THE FORMS OF ESCHATOLOGY

In the sphere of historical eschatology, distinctions should be made between the hopes of messianism, millennialism, and apocalypticism. Messianic hopes are directed toward a king of the end time who will lead the people of God, now suffering and oppressed, into a better historical future. In political and nativistic messianism, visions of the vengeance and of the equalizing justice on the side of the oppressed are aimed at political and religious leaders. Always at work in these instances are inner and often local historical expectations of a certain fulfillment of history before the end of history. Apocalypticism should thus be distinguished from this point of view. Apocalypticism upholds the view that God will intervene in history on the side of a faithful minority and that the intervention will be accompanied by sudden, cataclysmic events. According to this view, "this world" cannot bear the "justice of God." Therefore, against what is perceived as the perverted world, the followers of apocalyptic views hope for the creation of a new world on the basis of God's righteousness. If this hope is universal, it is, nevertheless, not generally represented by a people but rather by individual holy men or, perhaps, by an ascetic community. Millenarian hope is directed toward the 1,000-year Kingdom of Christ and of his own people, in which the ones who are suffering now will rule over their enemies.

Messianism. The term messiah, derived from the Hebrew word *mashiaḥ* ("anointed"), has been applied to a variety of redeemer figures, and many movements with a markedly eschatological or utopian-revolutionary character of message have been termed messianic. Though messianic movements have occurred throughout the world, they seem to be especially characteristic of the Jewish and Christian traditions. Hence, not only the word messiah but also other terms relating to the messianic type of phenomena are derived from biblical religion and from the history of Jewish and Christian beliefs—*e.g.,* "prophetic," "millenarian," and "chiliastic" movements—the last two terms referring to a 1,000-year reign of Christ and his saints before the final end of history. Moreover, the scientific study of messianic beliefs and movements—originating, as it did, in the Western theological and academic tradition—was directed mainly to phenomena occurring either in Christian history or in cultures exposed to Western colonial, missionary, and modernizing influences. These Western origins of messianic terms and concepts give discussions of the subject an almost unavoidable Judeo-Christian slant. Hence, many present-day sociologists and anthropologists have attempted to develop a more neutral terminology—*e.g.,* nativistic movements, religious movements of liberty and salvation, renewal movements, revitalization movements, crisis cults—but many of these terms emphasize incidental and adventitious aspects of the phenomenon and miss its essential features.

Apocalypticism. The term apocalypticism is generally restricted to eschatological views and movements in the West that focus on cryptic revelations about a sudden, dramatic, and cataclysmic intervention of God in history, the judgment of all men, the salvation of the faithful elect, and the eventual rule of the elect with God in a renewed heaven and earth. Western apocalypticism is based upon the archetypal apocalyptic work in the Judeo-Christian tradition, the Book of Daniel. Daniel is the only apocalyptic book to be admitted to the Old Testament canon, just as the Book of Revelation is the only apocalypse included in the canon of the New Testament. There are many noncanonical apocalyptic works from both Jewish and Christian authors, among them the three books of *Enoch,* the *Second Book of Esdras,* the *Ascension of Isaiah,* and the *Apocalypse of Peter.* All of the apocalyptic works written during the first efflorescence of millennialism, including the Book of Revelation, owe much of their shape and style to Daniel. This Old Testament book stands in the succession of the Jewish prophets and was apparently influenced by Iranian religious thought, such as the Zoroastrian concepts of the Last Judgment, the battle between good and evil involving both men and angels, and a punishment of fire for evildoers.

Millennialism. Millennialism (from the Latin word for 1,000) is a philosophy of history viewed from a Christian theological standpoint and a religious movement now associated with such modern Protestant sects as the Adventists, Jehovah's Witnesses, and certain segments of many Protestant denominations. There have been many millennial groups and individuals throughout church history. The term is derived from the imagery of the New Testament Book of Revelation (Rev. 20), in which the writer describes a vision of Satan being bound and thrown into a bottomless pit and of Christian martyrs being raised from the dead and reigning with Christ for 1,000 years. This 1,000-year period, known as the millennium, is viewed as a time during which man's yearnings for peace, freedom from evil, and the rule of righteousness upon earth are finally realized through the power of God.

As a branch of eschatology, millennialism is concerned with the earthly prospects of the human community. Not limiting itself to the prospects of the individual in this world and the next, millennialism attempts to answer in vivid imagery such questions as: What will be the final end of this world? Will mankind ever fulfill the agelong dream of dwelling in an earthly paradise or will all men be destroyed in a cataclysm of fire brought on by their own folly or God's judgment?

Millennialism is thus the cosmology (study of order) of eschatology, its chronology one of future events, comparable to the historical record of the past.

Millennialism is found within both Christian and other traditions. During the 20th century, anthropologists, historians, and sociologists explored the millennialist aspects of non-Western cultures, finding many striking similarities to the millennialism within the Judeo-Christian tradition. The millennial treatises produced by Jewish and Christian believers in the latter part of the Greco-Roman civilization—the Hellenistic period—particularly the books of Daniel and Revelation, provided the building materials from which the successive millennial structures were erected. In constant repetition the motifs, the leading characters, the symbols, and the chronologies of these works have arisen in the teaching of some prophet of the end of the world, each time taking on new significance from associations with contemporaneous events.

ESCHATOLOGICAL TERMINOLOGY

Eschatological language ordinarily uses two elements of style in conjunction with each other: the negation of the negative and the analogy of the future. Objective statements about the future that is not yet present are not possible in history. Such statements are possible only in the form of the negation of the negative in this life and this world. An example of this style may be found in Revelation 21:4: "And death shall be no more, neither shall there be mourning nor crying nor pain any more." Thus, the positive aspects of the eschatological future are circumscribed by the negative aspects of the present. If this future is to be meaningfully related to this life, however, this life, despite all its negativity, must also be presently capable of pointing toward or of foreshadowing the future life. Eschatological imagery and language, therefore, constantly use comparisons or analogies from everyday life (such as the several "the Kingdom of God is like . . ." analogies in the New Testament) and employ the hopes and anticipations of people and events from history in analogical ways.

The negation of the negative and the analogy of the future

A problem of eschatological language exists because of the necessary connection between negation and analogy. If negation is predominant, the tendency is towards apocalyptical and metaphysical dualism and towards mysticism. If analogy or foreshadowing of the future dominates eschatological language, the tendency is toward a one-sided belief in progress. In both cases the *novum* of eschatology becomes inexpressible. A hermeneutic (methodological interpretive principle) of eschatological traditions must verify the negation of the negative in face of the presently experienced negative and simultaneously seek the traces of the coming positive in the ambiguous history of liberations. There are always negative and positive signs of the future in history. "Where danger is, the saving also flourishes" (Hölderlin), but "where the saving draws near, danger also grows" (E. Bloch). Eschatology understands history as a growing crisis: the good provokes the evil, and the growing danger makes the action of redemption necessary. Authentic eschatology is neither world denying nor faith in progress but rather it can be seen as anticipation of freedom in the midst of slavery and of salvation in the midst of lostness and alienation.

ESCHATOLOGY IN NON-WESTERN RELIGIONS

Nonliterate cultures and nativistic movements. Eschatological motives must be understood in their religious and cultural contexts. According to the eschatological views of the people of the Andaman Islands, at the command of the god Puluga an earthquake will destroy the earth and the bridge of heaven; the souls and spirits of the dead will arise and be reunited. Then men will lead happy lives in power, without sickness, death, and marriage. Even the animals will appear again in their present form. The impatient spirits of the underworld are already now shaking the roots of the palm tree, which supports the earth, in order to bring about the end of this present world and the resurrection more quickly. The Semang pygmies in Malacca hold a somewhat different eschatological view:

The end of the order of the world

> In the beginning of the ages there was still absolutely nothing. Then Ja Pudeu [the highest being] blew with her mouth causing storms to rage over the earth. This was the means by which stars, water, trees, and everything came into being. At the end of the ages when all men shall have died, she will destroy everything with the same storms; a great flood will complete the destruction and the bones of men will swim together. Finally, the bones will rise.

The Australian aborigines claim that the end of the world will come when the moral world order legislated by the gods is no longer upright.

The Gabonese Pygmies in West Equatorial Africa believe that once Kmvum (the original man) lived with them faithfully. Then their guilt brought on the day of separation. He will come again, however, and bring back with him joy, abundance, and happiness. Among the Altaic Tatars of Central Asia there is an eschatological belief that Tengere Kaira Khan (the "graceful emperor of heaven"), who once lived on earth with men, will return at the end of the world in order to judge men according to their works. At his departure from earth he sent a mediator who remains faithful to him. At the end of the world the mediator will be victorious over evil. The Salish Indian tribes of the Pacific Northwest of North America believe that the creator god, before he vanished from the earth, promised the "elder" or "chief" his return at the end. When earth (a female figure) has become old, the coyote will return as the first sign of the world's end. This will be followed by the "chief" himself returning to earth. "After this there will no longer be a land of the spirits. All men will live together, the earth will receive her natural form and will live as a mother among her children. Then all things will be made right and happiness will reign."

Such mythical themes arranged around the origin–fall–return motif are experiencing a revival as primitive peoples suffer cultural shock in their encounters with Western civilization and Christianity. Many messianic movements of nonliterate cultures—even when antiwhite and anticolonialist—exhibit markedly Christian features both in the details of their symbolism as well as in their overall messianic ideology. Some messianic movements (*e.g.*, that of Simon Kimbangy in the former Belgian Congo from 1921, or that of Isaiah Shembe from 1911, among the South African Bantu, as also several movements in Brazil), in fact, appeared outwardly as Christian revivalist sects with an eschatological character. The movement of Simon Kimbangy has been admitted to the World Council of Churches as a member.

Nativistic messianic movements

A variety of names has been applied to these movements that emphasize various messianic characteristics. "Nativistic" movements expect salvation from a revival of native values and customs and a rejection of everything alien; *e.g.*, many of the North American Indian movements from the 17th century on, including the Pueblo Indian Revolt led by Popé in 1680; the anonymous Delaware prophet (1762) and Pontiac; the religious revival and revolt led by Tenskwatawa and Tecumseh in 1807; the Ghost Dance outbreaks of 1870 and subsequent years among Southwestern and Plains Indians. The messianic movements in Melanesia focussing on the arrival—in ships or airplanes—of "cargo" (*i.e.*, the coveted wealth and riches that symbolize power, well-being, and salvation) are referred to as cargo cults. Some anthropologists speak of "revitalization movements," whereas others emphasize the connection between acculturation and messianic movements. Since it is not acculturation as such that produces messianism but the crises and dislocations caused by certain forms of culture contact, many scholars prefer the more neutral and objective term "crisis cults." Since many movements are started or propagated by the activity and preaching of prophet-like leaders, they are also spoken of as "prophetic movements."

There is a tendency among modern anthropologists to consider primitive messianisms as forms of protonationalism in non-European and premodern societies. Though Christian influence and Christian symbols often play a major role in the crystallization of messianic ideologies, they are by no means their only source. The ideological starting point of a messianic movement can be supplied by native traditions and mythologies, by Christian ideas, or by motives that are born under the pressure of circumstance.

Religions of the East. Buddhism has four "noble truths"—the fact of human suffering, the understanding of the origin of suffering, the removal of the causes of suffering, and the path to the transcendence of suffering—and is partly therefore a religion of redemption. According to the Buddhist world view there is a macrocosm composed of innumerable worlds. In recompense for good and evil deeds, creatures are reborn in an unceasing process in the region that they deserve. Beyond all worlds is found Nirvāṇa ("bliss" or enlightenment), "the indescribable goal," whose attainment means redemption from the cycles of existence. Each one of the innumerable worlds passes through periods of destruction and recreation. There is no soul migration because there is no soul substance. Rather, each new existence is defined by *karman*, the deeds of the earlier existence. Only insofar as this is true can one speak of continuity in the reincarnations.

Buddhist and Hindu "eschatological" views

If anything at all in Buddhism can be considered as eschatological it is the yearning for an ultimate redemption from the cycle of rebirth and the final abolishment of the suffering bound up with it. In the religious community founded by the Buddha this yearning finds its peculiar way, the "eightfold path," on whose highest step occurs illumination or enlightenment. This illumination or "life beyond grief and woe" (Nirvāṇa) is a condition of eternal peace that cannot be conceived or described in ordinary terms. Rather, it comprehends even the worlds and their cycles themselves: "I do not know the end of suffering if we have not reached the end of the world" (Gautama, the Buddha).

According to Theravāda Buddhism, the individual believer must strive after redemption for himself through the exertion of his powers of being. At the centre of Mahāyāna Buddhism, on the contrary, stands the concept of the redemption of all living creatures which will occur through the sympathetic assistance of a redeemer figure.

In Hinduism the world of Brahman (the Universal Soul) is created by demiurges (creator beings) in egg-form and contains numerous zones and levels around the golden

world mountain Meru. It passes through a series of temporal cycles, and every temporal cycle ends with the destruction of the world followed by a new creation under a new demiurge. Man is now living in the dawn of the last, worst Kali *yuga* (age) of such a cycle. Through migration of the soul, man is drawn into the circle of the animals and plants. Rank and kind of birth are determined by the individual *karman* (acts and their consequences). The *karman*, which accompanies the soul and even has an effect on the destruction of the world, determines the following existence. After death, the soul returns to earth or it goes the "gods' way" (*devayāna*) of redemption. Redemption occurs when *karman* can no longer be produced, or it can happen through a divine act of grace that blots out the existing *karman*. Redemption is popularly viewed as entrance into the highest heaven of the god worshipped, where the redeemed one awaits a spiritual reflection of earthly joy. In modern Hinduism the soul that is identical with God is redeemed through a recognition of the organic wholeness that has vanished from consciousness because of the soul's imprisonment in matter. Self-recognition (*Ātman*) then leads to identity with the absolute being (Brahman). Redemption lies in the accomplishment, or rather recognition, of the *Ātman*-Brahman identity, for it already frees man from the chains of *karman* and *saṃsāra* (cycle of rebirths).

Buddhism and Hinduism do not have a historical eschatology. They rather emphasize an ultimate redemption from the cycles that have no beginning. Their redeemer figures are impersonal transparencies of the ultimate-universal.

Religions of ancient civilizations. In the religions of the West, there is a tendency to understand time as irreversible and therefore the end as occurring once and for all, as ultimate. The final judgment will be followed by the creation of a new and sacred world that is eternal. Origins of this kind of eschatological thinking are found in ancient Egypt in texts such as the "Shipwrecked Sailor" and the "Conversation between Atum and Osiris." The idea of an eschatological individual judgment of the dead is developed here in the strongest sense. Ancient Greek and Roman eschatological views depict a shadow life for the individual departed soul in Hades. Other than this focus on the individual, however, the cyclical concepts of periodic world destruction and world renewal are also found in these religions.

ESCHATOLOGY IN RELIGIONS OF THE WEST

Islām and Zoroastrianism. Islām is not a messianic religion and has no room for a saviour-messiah. Nevertheless, there gradually developed—probably under Christian influence—the notion of an eschatological restorer of the faith, identified as a descendant of the Prophet or as the returning 'Īsā (*i.e.*, Jesus). He is usually referred to as the *mahdī*; *i.e.*, the "[divinely] guided one." After the appearance of 'Īsā, the last judgment will begin: the good will enter paradise; the evil will fall into hell. Heaven and hell possess various goals and steps of recompense for good and evil. The time before the end is viewed pessimistically: God himself will abandon the godless world. Ka'bah (the great pilgrimage sanctuary of the Muslim world) will vanish, the copies of the Qur'ān will become empty paper, and its words will disappear from memory. Then the end will draw near.

In Sunnī (traditional) Islām the whole subject is one of folklore rather than of dogmatic theology, though all orthodox Muslims believe in the coming of a final restorer of the faith. In times of crisis and of political or religious ferment, mahdistic expectations have increased and have given rise to many self-styled *mahdī*s, the best known of all being Muḥammad Aḥmad, the Mahdī of Sudan, who raised a revolt against the Egyptian administration in 1881 and after several spectacular victories established the mahdist state that existed until defeated by the English military leader Kitchener at Omdurman (Sudan) in 1898. In the Islāmic Shī'ah sect (which holds a belief in the transference of spiritual leadership through the family of 'Alī, Muḥammad's cousin and son-in-law), the doctrine of the *mahdī* is an essential part of the creed. Among the Twelvers, the main Shī'ah group, the expected *mahdī* is believed to be the hidden 12th *imām*, or religious leader, who will reappear from his place of occultation. The notion of a *mahdī* also played a role in the foundation of new religions or Shī'ah sects—*e.g.*, the belief of the Druzes that the Egyptian caliph of the Fāṭimid dynasty al-Ḥākim (reigned 996–1021), who is thought to be the last prophet and divine incarnation, would return at the end of days (1,000 years after his appearance at the end of the 9th century AD) to establish his rule over the world. Other Islāmic-based messianic figures include the founder of the Indian Aḥmadiyah sect, Mirza Ghulam Ahmad, who in the late 19th century declared himself to be the Christ and the Mahdī; and the founder of the religion that subsequently became known as Bahā'ism, the Iranian Mirzā 'Alī Moḥammad of Shīrāz, who proclaimed himself in 1844 to be the Bāb ("gate") on the 1,000th anniversary of the disappearance of the 12th *imām*.

Zoroastrianism is a religion with a thoroughly eschatological orientation: for it world history is a battlefield on which the forces of light and good fight the powers of darkness and evil. Though the notion of a personal saviour figure is not essential to the Zoroastrian system, it did nevertheless arise. The Iranian prophet Zoroaster's ministry (6th century BC) is said to have opened the last of the history of the world's four periods of 3,000 years each. He is followed, at intervals of 1,000 years, by three "saviours," considered to be posthumous sons of Zoroaster. The last of these, the *soshyans* (or *saoshyant*), will appear at the end of days, and God will entrust to him the task of the final rehabilitation of the world and the resurrection of the dead.

Judaism. *In ancient times.* The real inception of historical eschatologies stems from the Old Testament. Israel's faith in God there is rooted in the historical experience of the Exodus, and because of this experience of liberation it contains a hope in the guidance and the promises of God in history. The basic structure of this faith is found in the law of promise and fulfillment. The eschatology of the Old Testament is grounded in the identity of faith in God and hope in the future (Gen. 12:1ff.). It has its beginning in the promise of a "good and broad land, a land flowing with milk and honey" (Ex. 3:8). In the Pentateuch (first five books of the Old Testament) the promise is broadened to the increase of people and possessions, the blessing of God and the victorious presence of God (Gen. 49:8–12; Num. 23; Deut. 33:13–17; Num. 23:21). The history of the occupation of the land of Canaan (Palestine) and the victory of the Hebrews is to be understood as "realistic hope." Through the experiences of Israel's own disobedience to the laws and the will of God and defeats at the hands of its enemies, the concept of the "day of the Lord," which is to bring salvation and victory, came into existence. The happiness of the establishment of the Kingdom by David in the 10th century BC led to a hope in the future Messiah (Anointed One) of God from the house of David (II Sam. 7).

Concept of the "day of the Lord"

In the midst of the political catastrophes of the 8th century BC, the great prophets took up the concept of the "day of the Lord" and proclaimed it as a day of judgment (Amos 5:18) over the disobedient people and also over all other peoples. It was through such a process that the day of the Lord concept became the bearer of eschatological hopes. Isaiah also viewed salvation in an eschatological light as happening only after the universal judgment (Isa. 4:3; 6:13; 11:11; 37:31) and combined it with the presence of a messianic mediator of salvation (7–12).

In spite of the political destruction of Israel (8th century BC) and Judah (6th century BC) the prophetic hope kept Israel alive religiously. This prophetic hope was aimed primarily at a comprehensive and total new creation, a new heart, a new covenant (Jer. 31; Ezek. 36; Isa. 41; Isa. 51). Through the vicarious atonement of the servant of God (either the people of Israel or a messianic figure) this hope was to include not only Israel but also the Gentile world (Isa. 42:6; 49:6). The future of Israel is thus bound up with the future of all peoples and of the whole earth.

In the incipient apocalyptic views of the prophet Daniel (2 and 7), hope is transcendent. His apocalyptic eschato-

logical hope expects the "Kingdom of the Son of man" following the consummation of evil in the fourth and final kingdom of the world. Since that time, hope in a Messiah and hope in a Son of man have been bound to a kind of eschatology that unites the fulfillment of the history of Israel with the end of world history.

In the face of a threat to the existence of the Jewish faith and temple worship, a group of zealots revolted in 168 BC against the occupation forces of the Seleucid monarch of Syria, Antiochus IV Epiphanes (*c.* 215–163/164 BC) and against those of their Jewish countrymen who favoured reducing Judaism to the level of a hellenized Oriental (Greek and Syrian) cult. The author of Daniel constructed his work to give aid and comfort to the rebel cause, particularly to assure them that God was aiding them, that the end of their struggles was in sight, and that a new golden age was dawning. In a vision reputed to be seen by King Nebuchadrezzar, he depicted a series of four world monarchies, represented in one passage by parts of a giant statue and in another by mythological beasts, each empire embodying evil to a greater extent than the last. Man's empires will end with the fourth kingdom, which is crushed by a "stone . . . cut out by no human hand," symbolizing the fact that neither its destruction nor the ensuing order are natural developments from forces latent in history. A figure called the Son of man, however, will institute a fifth, entirely righteous, just, and eternal kingdom.

Visions of the ages of the world

As in the Jewish prophetic tradition that preceded it, Daniel made predictions about the future, but, unlike the predictions of a prophet such as Jeremiah (late 7th to early 6th century BC), the outcome anticipated by the prophet was not the virtually inevitable product of antecedent forces but a total reversal of what might seem to be the likely outcome if God were not to intervene. The reversal of worldly expectations through a violent supernatural intervention in the course of history is one of the most characteristic features of apocalypticism and stands quite in contrast to the older prophetic style. Also essential to Daniel and subsequent apocalypticism is the immediacy of the message and the imminence of the deliverance that is promised—the promise of salvation now. Descriptions of this imminent salvation of cosmic proportions included vivid representations of historical figures who depicted the growth of evil and decline of goodness from past time down to the present, when all wickedness came into terrifying focus.

To these significant emphases in Daniel's apocalypse—the imminent and supernatural intervention of God in man's history and the reversal of the heretofore irresistible progress of evil and declension of good—might be added other characteristics that have proved to be influential. Numerology, mythological figures, and angelology, which have continued to play such a large part in millennial movements, were probably introduced as a result of the influence of Iranian thought. Other characteristics of Daniel, such as its pseudonymous authorship and the emphasis upon the esoteric, mysterious quality of the truths discussed, were probably due to the unique problems faced by the author in presenting these views to a 2nd-century-BC audience.

In Hellenistic Judaism. In the period of Seleucid (Syrian Greek dynasty ruling Palestine *c.* 200–165 BC) and later Roman and Byzantine (63 BC–AD 638) rule and oppression, the expectation of a personal messiah acquired increasing prominence and became the centre of a number of other eschatological concepts held by different groups in different combinations and with varying emphases. The Qumrān sect, a Jewish monastic group known in modern times for its preservation of the Dead Sea Scrolls, held a doctrine—found also in later Jewish sects—of a messianic pair: a priestly messiah of the House of Aaron (the brother of Moses) and a royal messiah of the House of David. This messianic detail, incidentally, shows that these "anointed ones" were not thought of as saviours—as in later Christian thought—but rather as ideal leaders presiding over an ideal, divinely-willed, and "messianic" socioreligious order. The "son of David" messianism, with its political implications, was overshadowed by apocalyptic notions of a more mystical and mythological character. Thus it was believed that a heavenly being called the "son of man" (the term is derived from Daniel 7:13) would descend to save his people (*e.g.*, as in the apocryphal books of Enoch). The messianic ferment of the period, attested by contemporary Jewish-Hellenistic literature, is also vividly reflected in the New Testament.

The destruction of the Temple at Jerusalem by the Romans (AD 70), exile, persecution, and suffering only intensified Jewish messianism, which continued to develop theoretically in theological and semimythological speculations and to express itself practically in messianic movements. In popular apocalyptic literature another messianic figure gained some prominence: the warrior-messiah of the House of Joseph (or Ephraim) who would precede the triumphant messiah of the House of David—but would himself fall in the battle against Gog and Magog, two legendary powers under Satan and opposed to the people of God (Ezek. 38:2; Rev. 20:8). The notion seems to have developed toward the end of the 2nd century, after the failure of the last revolt against the Romans (AD 132–135), led by Bar Kokhba, who was hailed as the messiah, but it is connected with a more basic notion of apocalyptic messianism; that is, the belief that the messianic advent is preceded by suffering and catastrophe. In some versions of apocalyptic messianism, the notion of a messianic age merges with that of an end of days and last judgment: the "new heaven and new earth" are ushered in amid destruction and catastrophe.

Apocalyptic messianism

In medieval and modern Judaism. Messianic faith tended to develop into mass enthusiasm, frequently fed by calculations based on the Book of Daniel and other biblical passages. Almost every generation had its messianic precursors and pretenders; *e.g.*, Abū ʿĪsā al-Iṣfahānī and his disciple Yudghan in 8th-century and David Alroy in 12th-century Persia; the propagandists of the messianic agitation in the Jewries of western Europe in the 11th and 12th centuries; and—perhaps the most notorious of all—the 17th-century pseudomessiah Shabbetai Tzevi (Sabbatai Zevi) of Smyrna. Belief in, and fervent expectation of, the messiah became firmly established tenets of Judaism and are included among the great Jewish medieval philosopher Maimonides' Thirteen Articles of Faith. There was much variety in the elaboration of the doctrine—from the early apocalyptic visionaries and later Kabbalistic (Jewish esoteric) mystics at one end of the scale to the rationalist theologians on the other. The latter (including Maimonides) emphasized the unmiraculous nature of the messianic age.

Modernist movements in Judaism tended to maintain the traditional faith in an ultimately redeemed world and a messianic future for mankind, without insisting on a personal messiah figure. Judaism undoubtedly owes its survival, to a considerable extent, to its steadfast faith in the messianic promise and future. Jewish messianism, in spite of its spiritual and mystical connotations, never relinquished its this-worldly orientation and its understanding of the messianic order in historical, social, and political terms. Hence, many writers consider the participation of Jews in so many secular reform, liberation, and revolutionary movements as a secularized version of traditional Jewish messianism. Similarly, the ideology of Zionism, as a movement for Jewish national emancipation and liberation, is not devoid of messianic features.

Individual eschatology emerges only on the periphery of the Old Testament. Amazingly, there were in Israel no known death cults and no vivid conceptions of life after death. The late expectation of the resurrection from the dead to judgment (Dan. 12:2) is not a yearning for salvation but hope in the victorious righteousness of God. Rabbinical messianism continued this same line of thought.

Christianity. *In the New Testament period.* The preaching and ministry of Jesus of Nazareth and the activities of his followers in the 1st century AD can be properly understood only in the context of contemporaneous Jewish eschatological beliefs. Though the precise nature of Jesus' beliefs about himself and about the nature of the "messianic" task that he attributed to himself are still a matter of scholarly controversy, there is little doubt that already at an early date his followers saw in him the promised "anointed one" (Greek *christos*, whence the En-

Messianic conceptions of Jesus

glish Christ) of the Lord, the son of David. This view is evident in the Gospel accounts that attempt to trace the ancestry of Jesus back to David, evidently for the purpose of legitimizing his messianic status. According to Luke 2:11 his messiahship was also proclaimed by angels at his birth. Jesus himself seems to have rejected the term—possibly because of its political implications—in favour of other eschatological titles (*e.g.,* the "Son of Man"), but the early community of his followers, believing, as they did, in his Resurrection after the crucifixion, evidently held this term to be expressive more than any other of the role and function that they attributed to their master and "Lord" (Greek *kyrios*). In due course the title ("Jesus, the Christ") became synonymous with the proper name, and the word Christ was used by believers as the name of the risen Jesus (*cf.* Gal. 1:6; Heb. 9:11).

With the adoption of the Greek word "Christ" by the church of the Gentiles (non-Jewish believers), the nationalist and political implications of the term "messiah" vanished altogether in Christianity, and the "Son of David" and the "Son of man" motifs, to which subsequently was added that of the "suffering Servant" (Isa. 52–53), could merge in a politically neutral and religiously original messianic conception. Subsequently, the doctrine of the messiahship of Jesus (*i.e.,* Christology) also had to take into account other features of evolving Christian dogma (the Messiah as the Son of God; the Trinity, of God the Father, Son, and Holy Spirit; the incarnation of the Word), and thus came to assert that Jesus as the Messiah, Saviour, and Redeemer was essentially divine. In due course the concept of salvation was radically spiritualized, and the Messiah, through his sacrificial death, was viewed as having delivered man from his bondage to sin and having restored him to communion with God. Meanwhile, Christians asserted that the present world order would provisionally continue until the Second Coming (the Parousia) of Christ in power and glory to judge the living and the dead.

The early Christians held this Second Coming to be imminent, but as time went on this particular expectation shifted to the eschatological horizon. In the centuries immediately following the writing of Daniel, the apocalyptic world view had significantly influenced Jewish culture; the audiences whom Jesus addressed were acquainted with it; and the early Christian Church embraced the apocalyptic world view. The Apostle Paul frequently expressed apocalyptic expectations (1 Thess. 4), and Mark 13, a passage often called "the little apocalypse," reflects the apocalyptic expectations of the Roman church at about AD 70.

The Christian Church in the 1st century wrestled with a difficult problem. Jesus had promised the inauguration of a new age, the Kingdom of God, and yet life proceeded after his death in much the same way that it had before his birth, with the exception that Christian believers suffered severe persecution for their faith. The primitive church solved this problem through the paradox of the Second Coming of Christ: Christ has come and is coming again. They believed that the new age had dawned but would not be fully revealed until Christ's Second Coming in glory.

The Book of Revelation

Like the Book of Daniel, the Revelation to John or Revelation was composed during a period of persecution. It was probably written during the last decade of the 1st century AD, and it reflects the persecutions beginning under the emperor Nero (AD 37–68), who seems to be portrayed as the Antichrist—the beast whose symbolic number is 666 (Rev. 13). After addressing letters to churches of Asia Minor, the author depicted his vision of a series of judgments—seven seals opened, seven trumpets blown, seven bowls poured out. The writer directed his attack against the Roman Empire, referred to cryptically as Babylon and as the great harlot. Christ was described as the executor of God's judgment, appearing not as the man Jesus but as an omnipotent king riding upon a white horse with eyes like a flame of fire and a mouth like a sharp sword "with which to smite the nations" (Rev. 19). In the Book of Revelation the assimilation of Jewish apocalypticism to Christianity was completed. Daniel's Son of man was replaced by Christ; many of the numerological formulas were repeated; and the dualistic universe of good and evil, Christ and Antichrist, was provided with a new and unforgettable set of characters. The essence of the apocalypse in Revelation remained what it had been in Daniel: the immediate, direct aid of God was to be momentarily expected, accomplishing the dramatic reversal of history that the believers' present desperate state demanded.

Apocalyptic symbolism

In the early church. During the first hundred years of Christian history, this form of millenarianism, or chiliasm (from the Greek word for 1,000), was commonly taught and accepted within the church. Persecution of the church was intermittent, however, and apocalyptic zeal flagged without the pressure of opposition. Christian missionaries succeeded in converting large numbers of Roman citizens, and some of the antagonism toward the empire was dissipated. The appeal of millenarian thought was further limited by its association with the heresy of Montanism. In characteristic apocalyptic fashion, Montanus, the founder of the movement, was fascinated with the idea of dividing past and future into units of prophetic calculation. In AD 156, according to the 4th-century Christian antiheretical writer Epiphanius, Montanus declared himself the prophet of a third testament, a new age of the Holy Spirit. Phrygia in Asia Minor became the centre of this ecstatic and ascetic movement whose leaders claimed divine inspiration for their visions and utterances, the main theme of which was the imminent Second Coming of Christ. This concept of the third age, the new day of the spirit of God, has been one of the most consistently repeated features of millenarian history, reappearing, for example, in Joachim of Fiore's philosophy of history during the 12th century, in views of the 17th-century Quakers, and in the apocalyptic speculations of the Seventh-day Adventists of the 19th and 20th centuries. When persecution of Christians was renewed late in the 2nd century, Montanism began to appeal outside Asia Minor and found converts throughout the Roman Empire, including Tertullian, a North African lawyer and theologian. The church survived this persecution, however, and Montanism was stigmatized as a heresy.

The age of the Holy Spirit

The influence of Greek thought upon Christian theology undermined the millenarian world view in another, possibly more significant, manner. In the theology of the great 3rd-century Alexandrian Christian thinker Origen, the focus was not upon the manifestation of the kingdom within this world but within the soul of the believer, a significant shift of interest away from the historical toward the metaphysical, or the spiritual. The association of apocalyptic millenarianism with the Montanist heresy, the growing influence of Greek thought upon Christian theology, and the conversion of Constantine the Great and adoption of Christianity as the favoured religion of the empire—all combined to discredit millenarianism for centuries.

The views of Augustine. In the new age of the church triumphant—*i.e.,* when Christianity became the accepted religion of the Roman Empire—Augustine (354–430), bishop of Hippo, gave definitive expression to the view that was to dominate Western civilization through the Reformation. In his *City of God,* a philosophy of history, Augustine viewed the world as eternally divided between the City of the World and the City of God. All men owe allegiance to one or the other of these cities and will ultimately share the fate of that community. The City of the World is ruled by Satan, the prince of this world. He and all who pay homage to his city will suffer eternal punishment. The City of God is represented in the church, and for the church God has ordained salvation from the persecution of the City of the World and eternal bliss in the courts of heaven. Much of this conception is reminiscent of the apocalyptic world view, but the views of Augustine also contained distinct differences.

The City of the World and the City of God

The dualism represented in apocalypticism is reflected with equal intensity in Augustine's two cities. Furthermore, Augustine remained as pessimistic as any millenarian about the future of the City of the World and the prospect of progress in this world. After his conversion to Christianity, Augustine, a former *bon vivant,* consistently favoured a world-denying and ascetic style of life. In fact, his disillusionment with worldly values was more thorough than that of the millenarians, for he rejected as carnal any expectations of a renewed and purified world that the believers could expect to enjoy. In this respect

he differed sharply with the apocalyptic tradition. The millenarian, in contrast to Augustine, had no quarrel with the world as such except that he had found it controlled by his enemies. The millenarian believed that when the imminently expected saviour had defeated these foes, the righteous would share in an earthly paradise, a land of physical, not spiritual, benefits.

The literalistic descriptions of the judgments that were predicted for the wicked and the bliss foretold for the righteous found in such apocalyptic works as the Book of Revelation were interpreted allegorically by Augustine. He expected that ultimately the history of this world would end, but for him the millennium had become a spiritual state into which the church collectively had entered at Pentecost—the time of the reception of the Holy Spirit by Christ's disciples soon after his Resurrection—and which the individual Christian might already enjoy through mystical communion with God. In contrast to the apocalypticist's focus upon the contemporary world, Augustine, though just as influenced by his own cultural milieu, responded with a millennial eschatology that seemed almost oblivious of time. As far as the struggle with evil in this world is concerned, Augustine surrendered and abandoned the field. No imminent supernatural intervention in history was expected, and no dramatic reversal of the tide of battle was anticipated. Augustine taught what has been referred to as "realized" eschatology. For him the battle had already been fought on the spiritual ground that really mattered. God had triumphed. Satan has been reduced to lordship in this world. In the present age the City of the World and the City of God have been forced to coexist. Eventually, even that small patrimony that Satan claimed would be taken from him, and God would become triumphant.

Realized eschatology

Medieval and Reformation millennialism. Augustine's allegorical millennialism became the official doctrine of the church, and apocalypticism went underground. During the late Middle Ages and the Reformation, millenarian views were frequently voiced, most often by rebels and radicals. The extreme wing of the Bohemian Hussite movement, known as the Taborites, sought to establish the Kingdom of God by force of arms. The left-wing Protestant Anabaptists as well as the Bohemian and Moravian Brethren were millenarians. The great Peasants' War in Germany (1524–25), in which the radical reformer Thomas Müntzer and the radical Zwickau prophets took a leading part, and the Anabaptist "Kingdom of God" in the German city of Münster (1534–35)—ruled over by the fanatical John of Leiden—are examples of millenarian-apocalyptic movements or of social movements with a messianic dimension.

In England, the Independents (those who separated themselves from the Church of England) thought of ushering in the Kingdom of God, and groups such as the Fifth Monarchy Men believed that revolution was necessary to prepare the way for the reign of Christ and his saints. The revolutionary Puritan leader Oliver Cromwell's (1599–1658) sober common sense and his dissolution of the so-called Parliament of Saints prevented apocalyptic enthusiasm from dominating the Commonwealth. The millenarian element also was strong in 17th- and 18th-century German Pietism, and it played a major role in the doctrines of many sects that arose in the 19th century in the United States and Great Britain (*e.g.*, Irvingites, Mormons, Adventists, Jehovah's Witnesses, Christadelphians, and others). Many of these sects, however, are more correctly described as entertaining messianic expectations than as actual messianic movements.

Apart from these dissidents, the doctrine of Augustine remained unchallenged until the 17th century. The Protestant Reformers of the Lutheran, Calvinist, and Anglican traditions were not apocalypticists but remained firmly attached to the views of Augustine, for whose theology they felt a particular affinity. Many of the allusions of the Book of Revelation were viewed in a distinctly Protestant perspective by the Reformers—the allusions to Rome as the great harlot and as Babylon being transferred to the Roman church, and the pope being identified as the beast. Each of the three main Protestant traditions of 16th-century Europe, however, found support from the secular authorities in Saxony, Switzerland, and England and remained in the same position vis-à-vis the state as had the medieval church. The apocalyptics within medieval Christendom, as well as in the 16th-century Reformation, were those who believed that their only help was the Lord and for whom persecution was a reality and destruction an imminent threat.

The Augustinian millennial world view, though it survived the Reformation, did not survive the intellectual revolution of the 17th century. Behind the development of science lay a profound reorientation of Western thought that involved, in the first place, the rehabilitation of nature. A part of Augustine's rejection of the world stemmed from the frustration felt by his generation in attempting to cope with the natural and social history of its time. By 1600 Europeans had gained confidence in their own abilities. Such philosophers as Francis Bacon announced the dawn of a new day and attacked the Augustinian reluctance to see anything but the work of the devil in attempts to control or understand natural processes. Secondly, European intellectuals were becoming far more interested in measurement and quantification. Allegory fell into disrepute when the medieval interpretation of the nature of the heavenly bodies was proved to be erroneous by the facts discovered by the use of the telescope. A new concern with calculation and literalism spread to biblical scholarship and resulted in the creation of the third type of millennialism found in the Christian tradition—progressive millennialism.

Progressive millennialism

Early progressive millennialism. Joseph Mead (Mede), a 17th-century Anglican biblical scholar, was the pioneer in the movement. Ignoring the allegorical interpretation long associated with the book, Mead took a fresh look at the text of Revelation. He concluded that the Scriptures held the promise of a literal Kingdom of God. The work of redemption, he concluded, would be completed within human history on the stage of this world. The Book of Revelation itself seemed to contain a historical record of the progress of that Kingdom, and scholars other than Mead soon were speculating where in the prophetic timetable the modern millennialist might locate himself. Thus far, progressive millennialism appeared to be identical with the apocalyptic millenarianism of the early church, but there the similarity ended. The Kingdom would not be brought into being through any dramatic reversal of the historical process, nor did the progressive millennialists believe that the Second Advent of Christ would occur in order to rescue them from destruction. History did not need reversing for these early Enlightenment Christians (those who emphasized reason). They thought of the record of the past as the story of victory over evil and the conquest of Satan. They rejected the fundamental assumptions of the apocalypticist—*i.e.*, that victory would be snatched from the jaws of defeat only by a miraculous deliverance. For them it seemed that the progress of history had been continuously upward, that the Kingdom of God was coming ever closer, and that it would arrive, not without struggle, but on the basis of the same kind of effort that had always triumphed in the past.

Millenialist optimism

In the 18th century the teachings of the progressive millennialists became dominant in many Protestant churches. The Anglican polemicist and commentator Daniel Whitby (1638–1726), in his *Paraphrase and Commentary on the New Testament* (1703), provided such convincing support to the position that he has often been credited with creating it. In America interest in the millennium had not been lacking among Puritan scholars, but it was the great revivalist Jonathan Edwards (1703–58) who first adopted progressive millennialism, giving detailed exposition of his views in his uncompleted *History of the Work of Redemption.* Edwards saw significance for millennialism in the discovery and settlement of the New World, and he anticipated the establishment of Christ's kingdom sometime near the end of the 20th century.

Later progressive millennialism. The association of the millennium with the role of the United States proved to be a volatile 19th-century mixture in the hands of Protestant ministers, and for much of that period millennialism

fed the fires of nationalism and Manifest Destiny. In a typical utterance, a leading Presbyterian minister of the 1840s, Samuel H. Cox, told an English audience that, "in America, the state of society is without parallel in universal history. . . . I really believe that God has got America within anchorage, and that upon that arena, He intends to display his prodigies for the millennium." The late 19th-century movement known as the Social Gospel, dedicated as it was to establishing the Kingdom of God here and now, manifested most clearly the continuing influence of progressive millennialism.

Because the advocates of optimistic millennialism were confident of the ultimate triumph of their cause, it must not be assumed that they took evil lightly. They thought of God's Kingdom as advancing, as Jonathan Edwards argued, but not without destruction. Though they were not apocalyptic, their view of history included the cataclysmic. In the American Civil War, for example, the antislavery writer Julia Ward Howe, in "The Battle Hymn of the Republic," described God's truth as "marching on." In Pres. Woodrow Wilson's crusade to make the world "safe for democracy" by the entry of the United States into World War I (1914–18), one can see the same idea barely disguised. According to the progressive millennialists, Christ's Second Advent would occur at the close of the millennium as its crowning event, and, as a result, their position has frequently been called postmillennialism.

ESCHATOLOGY IN MODERN TIMES

Western civilization, even in its modern secularized forms, is heir to a long tradition of Christian patterns of thought and sensibility. Thus, it is not surprising that many movements of social reform as well as ideologies regarding an ideal future should bear traces—conscious or unconscious—of Christian influence. Both the 18th- and 19th-century Enlightenment and the Romantic versions of the idea of the progress of humanity to an ideal state of peace and harmony betray their descent from messianic-millenarian beliefs. The 18th-century German philosopher Immanuel Kant, when speaking of the ideal state of eternal peace, describes this concept as a "philosophical chiliasm." The indebtedness of presocialist, utopian thinkers—such as the French social reformer Henri de Saint-Simon, the English reformer Robert Owen, and the French reformer Charles Fourier—to Christian millenarianism was recognized by Karl Marx and Friedrich Engels, who, in their *Communist Manifesto* (1848), contemptuously referred to the utopias of these writers as "duodecimo editions of the New Jerusalem." Some early socialist movements, including Christian socialism, exhibited messianic features.

Analogies of Marxist Communism with Christian messianism

Marxist Communism, in spite of its explicit atheism and dogmatic materialism, has a markedly messianic structure and message.

Some of the analogies between Marxism and traditional Christian eschatology have been described, in a slightly ironical vein, by the English philosopher Bertrand Russell, who contends that Marx adapted the Jewish messianic pattern of history to socialism in the same way that the philosopher-theologian St. Augustine (AD 354–430) adapted it to Christianity. According to Russell, the materialistic dialectic that governs historical development corresponds—in the Marxist scheme—to the biblical God, the proletariat to the elect, the Communist Party to the church, the revolution to the Second Coming, and the Communist commonwealth to the millennium.

Whether or not Socialism and Communism, as well as certain national liberation movements, are described as secularized messianism, pseudomessianism, "substitute" messianism, and the like, is partly a matter of semantics, partly an attempt to use evaluative instead of descriptive language. The differences between secular ideologies and traditional messianic expectations are obvious. The similarities are founded on actual historic contacts and derivation (as in the history of reform and revolutionary movements in the West as well as of liberation movements in countries colonized by the West), and also on the fact that they are variations of the same social dynamisms and of a basic myth, expressing in powerful imagery certain elemental human experiences and aspirations.

Since the exegetical (interpretive) works of Johannes Weiss and Albert Schweitzer around 1900 (school of "consistent eschatology") and dialectic theology (Karl Barth, Rudolf Bultmann), eschatology has again become a principal theme of academic Christian theology. The crises in the Western countries have led to a renewed activization of eschatological hopes. In church struggles, this was expressed in terms of distinctions between Christianity as a state religion and congregations with eschatological orientations. On its margins, Western civilization contains a series of mystical and apocalyptical "anticultures." Initial attempts to combine eschatology and philosophy, hope, and social practice, and thus overcome the difference between the church and the sects, as well as the church and the modern age, are found in Ernst Bloch's philosophy of hope (*Das Prinzip der Hoffnung*, 1959), the writings of P. Teilhard de Chardin, and in the "theology of hope" (J. Moltmann, W. Pannenberg, H. Cox, L. Dewart, etc.). Eschatology is one of the main focuses of Christianity. Therefore, its most important theological decisions about theory and practice must take eschatological concerns seriously, as seriously as the many and varied revolutionary groups (both religious and secular) have done in the 20th century. (J.D.M./R.J.Z.W./E.R.S./Ed.)

Angels and demons

Throughout the history of religions, varying kinds and degrees of beliefs have existed in various spiritual beings, powers, and principles that mediate between the realm of the sacred or holy—*i.e.*, the transcendent realm—and the profane realm of time, space, and cause and effect. Such spiritual beings when regarded as benevolent are usually called angels in Western religions; those viewed as malevolent are termed demons. In other religions—Eastern, ancient, and those of nonliterate cultures—such intermediate beings are less categorical, for they may be benevolent in some circumstances and malevolent in others.

NATURE AND SIGNIFICANCE

The meaning of angels in functional terms

Angels. The term angel, which is derived from the Greek word *angelos,* is the equivalent of the Hebrew word *mal'akh,* meaning "messenger." The literal meaning of the word angel thus points more toward the function or status of such beings in a cosmic hierarchy rather than toward connotations of essence or nature, which have been prominent in popular piety, especially in Western religions. Thus, angels have their significance primarily in what they do rather than in what they are. Whatever essence or inherent nature they possess is in terms of their relationship to their source (God, or the ultimate being). Because of the Western iconography (the system of image symbols) of angels, however, they have been granted essential identities that often surpass their functional relationships to the sacred or holy and their performative relationships to the profane world. In other words, popular piety, feeding on graphic and symbolic representations of angels, has to some extent posited semidivine or even divine status to angelic figures. Though such occurrences are not usually sanctioned doctrinally or theologically, some angelic figures, such as Mithra (a Persian god who in Zoroastrianism became an angelic mediator between heaven and earth and judge and preserver of the created world), have achieved semidivine or divine status with their own cults.

In Zoroastrianism there was a belief in the *amesha spentas,* or the holy or bounteous immortals, who were functional aspects or entities of Ahura Mazdā, the Wise Lord. One of the *amesha spentas,* Vohu Manah (Good Mind), revealed to the Iranian prophet Zoroaster (6th century BC) the true God, his nature, and a kind of ethical covenant, which man may accept and obey or reject and disobey. In a similar manner, about 1,200 years later, the angel Gabriel (Man of God) revealed to the Arabian prophet Muḥammad (5th–6th century AD) the Qur'ān (the Islāmic scriptures) and the true God (Allāh), his oneness, and the ethical and cultic requirements of Islām. The epithets used to describe Gabriel, the messenger of God—"the spirit of holiness" and "the faithful spirit"—are similar to those

applied to the *amesha spentas* of Zoroastrianism and the third Person of the Trinity (Father, Son, and Holy Spirit) in Christianity. In these monotheistic religions (though Zoroastrianism later became dualistic) as also in Judaism, the functional characteristics of angels are more clearly enunciated than their ontological (or nature of Being) characteristics—except in the many instances in which popular piety and legend have glossed over the functional aspects.

Various religions, including those of nonliterate cultures, have beliefs in intermediary beings between the sacred and profane realms, but the belief is most fully elaborated in religions of the West.

Changes and variations in the meaning of the term demon

Demons. The term demon is derived from the Greek word *daimōn,* which means a "supernatural being" or "spirit." Though it has commonly been associated with an evil or malevolent spirit, the term originally meant a spiritual being that influenced a person's character. An *agathos daimōn* ("good spirit"), for example, was benevolent in its relationship to men. The Greek philosopher Socrates, for example, spoke of his *daimōn* as a spirit that inspired him to seek and speak the truth. The term gradually was applied to the lesser spirits of the supernatural realm who exerted pressures on men to perform actions that were not conducive to their well-being. The dominant interpretation has been weighted in favour of malevolence and that which forbodes evil, misfortune, and mischief.

In religions of nonliterate peoples, spiritual beings may be viewed as either malevolent or benevolent according to the circumstances facing the individual or community. Thus, the usual classification that places demons among malevolent beings is not totally applicable in reference to these religions.

The positions of spiritual beings or entities viewed as benevolent or malevolent may, in the course of time be reversed. Such has been the case in the ancient Indo-Iranian religion, from which evolved early Zoroastrianism and the early Hinduism reflected in the Vedas (ancient Aryan hymns). In Zoroastrianism the *daevas* were viewed as malevolent beings, but their counterparts, the *devas* in ancient Hinduism, were viewed as gods. The *ahuras* of Zoroastrianism were good "lords," but in Hinduism their counterparts, the *asuras,* were transformed into evil lords. In a similar manner, Satan, the prosecutor of men in the court of God's justice in the Old Testament book of Job, became the chief antagonist of Christ in Christianity and of man in Islām. Many similar transformations indicate that the sharp distinctions made between angels as benevolent and demons as malevolent may be too simplistic, however helpful such designations may be as indicators of the general functions of such spiritual beings.

CELESTIAL AND NONCELESTIAL FORMS: RELATIONSHIPS OF BELIEFS IN ANGELS AND DEMONS TO VIEWS OF THE COSMOS

Man's concern with ways of relating to the cosmos

Because man is a being much concerned with boundaries—*i.e.,* what makes him different from other animate beings, what makes his community (and thus his world) different from other communities (and other worlds)—his view of the cosmos has influenced his understanding of what are called angels and demons. The cosmos may be viewed as monistic, as in Hinduism, in which the cosmos is regarded as wholly sacred or as participating in a single divine principle (Brahman, or Being itself). The cosmos may also be viewed as dualistic, as in Gnosticism (an esoteric religious dualistic belief system, often regarded as a Christian heretical movement, that flourished in the Greco-Roman world in the 1st and 2nd centuries AD), in which the world of matter was generally regarded as evil and the realm of the spirit as good. A third view of the cosmos, generally found in the monotheistic religions of Judaism, Zoroastrianism, Christianity, and Islām, centred on a tripartite universe: celestial, terrestrial, and subterrestrial. This third view has influenced Western man's concepts of angels and demons as well as his scientific and metaphysical concepts.

Relationship to views of a tripartite cosmos. In the biblical, Hellenistic (Greco-Roman cultural), and Islāmic worlds of thought, the terrestrial realm was a world in which man was limited by the factors of time, space, and cause and effect. The celestial realm, generally composed of seven heavens or spheres dominated by the seven then-known planets, was the realm of the divine and the spiritual. The subterrestrial realm was the area of chaos and the spiritual powers of darkness. At the highest level of the celestial sphere was the ultimate of the sacred or holy: *e.g.,* Yahweh, the God of Judaism, whose name was so holy it should not even be spoken; Bythos, the unknowable beginning beyond beginnings of Gnosticism; the heavenly Father of Christianity, known through his Logos (the divine Word, or Reason, Jesus Christ); and Allāh, the powerful, the almighty, and the sublime God of Islām.

The role of angels as celestial beings

In order to reveal the purpose and destiny of man—the highest being of the terrestrial realm—the ultimate of the celestial sphere enabled man, according to such views, to come to a knowledge of who he is, what is his origin, and what is his destiny through celestial messengers—angels. The message, or revelation, was usually focussed on the identity of the source of the revelation—*i.e.,* the ultimate being—and on the destiny of man according to his response. Because of a cosmic rift in the heavenly sphere prior to the creation of the world or the announcement of the revelation, angels, depending on their relationship to the Creator, might attempt to deceive man with a false revelation or to reveal the truth about man's true nature (or identity), origin, and destiny. Angels who attempted to pervert the message of the ultimate celestial being in order to confuse man's understanding of his present boundary situation as a terrestrial being or his destiny as a supraterrestrial being—though not always termed demons—are malevolent in function. Included among such malevolent angels are the devil of Christianity and Judaism or Iblīs (the Devil) of Islām, who, in the form of a serpent in the biblical story of the Garden of Eden—according to later interpretations of the story—attempted to disrupt man's understanding of his creaturely boundaries, or limitations. He did this by tempting man to eat the fruit of the tree of knowledge of good and evil so that he might become like God (or the divine beings of the heavenly court). In Zoroastrianism, the Evil Spirit (Angra Mainyu, later Ahriman) attempted—through subservient spirits such as Evil Mind, the Lie, and Pride—to deceive terrestrial man so that he would choose a destiny that was subterrestrial—punishment in a chasm of fire.

In the aftermath of the 16th-century Copernican revolution (based on the theories of the Polish astronomer Copernicus), in which man's view of the cosmos was radically altered—*i.e.,* the Earth was no longer seen as the centre of the cosmos but, instead, merely as a planet of a solar system that is a very small part of a galaxy in an apparently infinite universe—the concepts of angels and demons no longer seemed appropriate. The tripartite cosmos—heaven above, Earth in the middle, and hell below—appeared to be an anachronism.

Influence of psychology on angelology and demonology

With the emergence of modern Western psychology and psychoanalytical studies in the 19th and 20th centuries, however, the underlying principles of beliefs in angels and demons have taken on new meanings. Many Christian theologians have found some of the concepts of psychoanalysis helpful in reinterpreting the meanings underlying primitive and traditional beliefs in angels and demons. The tripartite cosmos was re-mythologized into a tripartite structure of the personality—the superego (the restrictive social regulations that enable man to live as a social being), the ego (the conscious aspects of man), and the id, or libido (a "seething, boiling cauldron of desire that seeks to erupt from beneath the threshold of consciousness"). Thus, demons—according to this reinterpretation—might well be redefined as projections of the unregulated drives of man that force him to act only according to his own selfish desires, taking no account of their effects on other persons. From a social point of view, demons might also be defined as the environmental and hereditary forces that cause man to act, think, and speak in ways that are contrary to the well-being of himself and his community. A modern French writer, Denis de Rougemont, has maintained in his book *The Devil's Share* that the devil and the demonic forces that plague the modern world can be

well documented in modern man's return to barbarism and man's inhumanity to man. In the 2nd century AD, Clement of Alexandria, a Christian philosophical theologian, pointed toward a psychological interpretation of demonic forces by stating that man was often captivated by the inner appetitive drives of his passions and bodily desires. The Freudian "myth" of the human personality and other psychological studies have thus initiated a new dimension in the study of angels and demons. Medieval iconography, which graphically depicted angels and demons as hybrid creatures that often defied even the most vivid imaginations of the persons who viewed them, has been supplanted by psychological, psychoanalytical, and modern mythological symbolism coupled with theological reflection.

Relationship to views of a dualistic cosmos. In religious traditions that have viewed the cosmos in a dualistic fashion, such as Gnosticism, angels were believed to be celestial beings who controlled certain spheres through which a soul was to pass as it freed itself from the shackles of its material existence. Knowledge of these angels and their names was a necessary prerequisite for achieving eventual union with the ultimate spiritual reality. Included among various lists of the seven angels ruling the seven planetary spheres are Gabriel, Adonai (Lord), Aariel (lion of God), and others. The angel of the creation of the world of matter, Yahweh (sometimes called the Demiurge, the Creator), was evil, in the Gnostic view, not only because he was the Creator but also because he tried to keep spiritual men from knowing their true origin, nature, and destiny.

Angels as aids or hindrances to man's goal of salvation

Manichaeism, a dualistic religion founded in the 3rd century AD by Mani, an Iranian prophet, like Gnosticism divided the world into two spheres—Goodness (Light) and Evil (Darkness). These two principles are mixed in the world of matter, and the object of salvation is to unmix the material and the spiritual so that one may achieve a state of absolute goodness. Highest in the celestial hierarchy are the 12 light diadems of the Father of Greatness and the Twelve Aeons, the "firstborn"—angelic figures that are divided into groups of threes, surrounding the Supreme Being in the four quarters of the heavens. Because the Devil, the Prince of Darkness, desires the advantages of the Kingdom of Light, in an ensuing battle between the celestial forces Light and Darkness are mixed, and the world of matter and spirit is created. Unaware of his spiritual nature and constantly tempted by the demons of the Prince of Darkness, man is eventually led to understand his true nature through the activity of angelic beings called the Friends of the Lights and the Living Spirit and his five helpers: Holder of Splendour, King of Honour, Light of Man, King of Glory, and Supporter.

Relationship to views of a monistic cosmos. Those who view the cosmos as basically monistic—such as Hinduism, Jainism, and Buddhism—generally have no belief in angels, who function mainly as revealers of the truth. This function is performed by other beings, such as *avatāras* (incarnations of the gods) in Hinduism, *tīrthaṅkaras* (saints or prophets) in Jainism, or *bodhisattvas* (Buddhas-to-be) in Buddhism. Because such personages generally are viewed more in terms of exemplifiers of the holy life than as conduits of a revelation (except in the case of several *avatāras* and *bodhisattvas*), they are not to be regarded in terms of the typical Western conceptions of angelic beings. These religions do, however, have widespread beliefs in demons.

Belief in demons as common to all religious or mythological views about the cosmos. Belief in demons is not connected with any particular view of the cosmos. Demons have a very wide geographical and lengthy historical role as spiritual beings influencing man in his relationship to the sacred or holy. They may be semihuman, nonhuman, or ghostly human beings who, for various reasons, generally attempt to coerce man into not attaining his higher spiritual aspirations or not performing activities necessary for his well-being in the normal course of living. The ancient Assyrian demon *rabiṣu* apparently is a classic prototype of a supernatural being that instilled such a fear in men that their hair literally raised from their bodies when confronted with knowledge of the *rabiṣu*'s presence.

Powers of demons in the West

In 17th-century Europe, various demons were cataloged according to their powers to entice men to indulge in what were called their basic instincts or desires. Included in such lists were nightmare demons, demons formed from the semen of copulation, and demons who deceived persons into believing that they could perform transvections (nocturnal flights to sites of sabbats, alleged rites of witchcraft). According to some authorities in the 20th century (as well as early Christian polemicists), the alleged demons noted by the prevailing religions of the world are the former gods or spiritual beings that succumbed to or were overpowered by the dominant doctrinal views of a conquering people. Thus, the Teutonic, Slavic, Celtic, or Roman gods either were reduced to demonic antagonists of Christ, his saints, or his angels or were absorbed by the cults of Christian saint figures. Followers of the ancient but no longer influential deities were often subjected to persecution as advocates of witchcraft, especially in Christian Europe (see also OCCULTISM: *Witchcraft*).

TYPES OF ANGELS AND DEMONS

Angels and demons, as noted earlier, have been categorized as benevolent, malevolent, or ambivalent or neutral beings that mediate between the sacred and profane realms.

Benevolent beings. Benevolent beings, usually angels but sometimes ghosts of ancestors or other spiritual beings that have been placated by sacrifices or other rituals, assist man in achieving a proper rapport with God, other spiritual beings, or man's life situations. Angels, for example, not only act as revealers of divine truths, but they also are believed to be efficacious in helping man to attain salvation or special graces or favours. Their primary function is to praise and serve God and do his will. This is true of angels in both Christianity and Zoroastrianism, as well as in Judaism and Islām. As functional extensions of the divine will, they sometimes intervene in human affairs by rewarding the faithful and punishing the unjust or by saving the weak, who are in need of help, and destroying the wicked, who unjustly persecute their fellow creatures. In the intertestamental book of Tobit (an apocryphal, or "hidden," book that is not accepted as canonical by Jews and Protestants), the archangel Raphael (God Heals), for example, helps the hero Tobias, the son of Tobit, on a journey and also reveals to him magic formulas to cure his father's blindness and to counteract the power of the demon Asmodeus.

Functions of angels

Angels also have been described as participants in the creation and the providential continuance of the cosmos. Clement of Alexandria, influenced by Hellenistic cosmology, stated that they functioned as the movers of the stars and controlled the four elements—earth, air, fire, and water. Many angels are believed to be guardians over individuals and nations. The view that there are guardian angels watching over children has been a significant belief in the popular piety of Roman Catholicism. Angels are also regarded as the conductors of the souls of the dead to the supraterrestrial world. In the procreation of men, angels are believed to perform various services. This is especially noticeable in the instances of angels announcing the births of divine figures or special religious personages, such as Jesus and John the Baptist in the New Testament.

Though the function of angels is of primary significance, theological reflection and popular piety have placed much emphasis on the nature of angels. In early Judaism angels were conceived as beings in human form: the angel who wrestled with the patriarch Jacob, as recorded in the book of Genesis, was in the form of a man. In Judaism of the Hellenistic period (3rd century BC to 3rd century AD), however, angels were viewed as noncorporeal spiritual beings who appeared to man in an apparitional fashion. Their spiritual nature had been emphasized earlier by Old Testament prophets, such as Ezekiel and Isaiah, in their visionary descriptions. The cherubim and seraphim, two superior orders of angels, are described as winged creatures that guard the throne of God. The use of wings attached to various beings symbolizes their invisible and spiritual nature, a practice that can be traced back to the ancient Egyptians, who represented the battling sun-god Horus of Edfu as a winged disk. In Christian iconography

the spiritual nature of angels has been almost universally represented—until the 20th century—by winged human figures. Their spirituality and, therefore, their noncorporeality led to various kinds of speculation among theologians and common people about the nature of the appearances of angels, which has been recorded in both Scripture and legends based on popular piety. Some theologians, such as Augustine in the 4th and 5th centuries, stated that angels, who have ethereal bodies, may be able to assume material bodies. This problem, however, has not been solved to the satisfaction of later theologians.

Malevolent beings. Malevolent beings—demons, fallen angels, ghosts, goblins, evil spirits in nature, hybrid creatures, the *daevas* of Zoroastrianism, the *nārakas* (creatures of hell) of Jainism, the *oni* (attendants of the gods of the underworld) in Japanese religions, and other such beings—hinder man in achieving a proper relation with God, the spiritual realm, or man's life situations. Some angels are believed to have fallen from a position of proximity to God—such as Lucifer (after his fall called Satan by early Church Fathers) in Judaism, Christianity, and Islām—because of pride or for attempts to usurp the position of the Supreme Being. In their fallen condition they then attempt to keep man from gaining a right relationship with God by provoking men to sin. Some medieval scholars of demonology ascribed to a hierarchy of seven archdemons the seven deadly sins: Lucifer (Pride); Mammon (Avarice); Asmodeus (Lechery); Satan (Anger); Beelzebub (Gluttony); Leviathan (Envy); and Belphegor (Sloth). Besides tempting men to sin, the fallen angels, or devils, were believed to cause various types of calamities, both natural and accidental. Like the demons and evil spirits of nature in primitive religions, the fallen angels were viewed as the agents of famine, disease, war, earthquakes, accidental deaths, and various mental or emotional disorders. Persons afflicted with mental diseases were considered to be "demon possessed."

The function of demons

Though the functions of demonic figures, like those of fallen angels, is of major significance, the nature of demons has been of concern to theologians and persons infused with popular piety. Like angels, demons are regarded as spiritual, noncorporeal beings, but they have been depicted in religious iconography as hybrid creatures with horrifying characteristics or as caricatures of idols of an opposing religion. In the early church, for example, there was a belief that pagan idols were inhabited by demons. The horrifying aspects of demons have been represented in the woodcuts of medieval and Reformation artists and in the masks of shamans, medicine men, and priests of primitive religions—either to frighten the believer into behaving according to accepted norms or to ward off ritualistically the power of the demonic forces loose in the terrestrial or profane realm.

Ambivalent or neutral beings. Ambivalent or neutral spiritual beings are usually not found in Western religions, which usually divide the inhabitants of the cosmos into those who are either allied with or in opposition to the Supreme Being. Islām, however, classifies spiritual beings into angels (*malā'ikah*), demons (*shāyaṭīn*), and *djinni*, or genies. This last category includes spiritual beings that might be either benevolent or malevolent. According to legend, the *djinni* were created out of fire 2,000 years before the creation of Adam, the first man. Capable of both visibility and invisibility, a *djinni* could assume various forms—either animal or human—and could be either a help or a hindrance to man. By cunning, a superior use of intellect, or magic, a man might be able to manipulate a *djinni* for his own benefit.

The role of the *djinni*

Various minor nature spirits—such as the spirits of water, fire, mountains, winds, and other spirits recognized in primitive religions—are generally neutral, but, in order to keep them that way or to make them beneficial to man, proper sacrifices and rituals must be performed.

VARIETIES OF ANGELS AND DEMONS IN THE RELIGIONS OF THE WORLD

Intermediate beings between the sacred and profane realms assume various forms in the religions of the world: celestial and atmospheric beings; devils, demons, and evil spirits; ghosts, ghouls, and goblins; and nature spirits and fairies.

In Zoroastrianism, Judaism, Christianity, and Islām. In the Western religions, which are monotheistic and view the cosmos as a tripartite universe, angels and demons are generally conceived as celestial or atmospheric spirits. In the popular piety of these religions, however, there is a widespread belief in ghosts, ghouls, goblins, demons, and evil spirits that influence man in his terrestrial condition and activities. The celestial beings may be either benevolent or malevolent, depending on their own relationship to the Supreme Being. On the other hand, the demons and evil spirits that generally influence man in his role as a terrestrial being (rather than in his destiny as a supraterrestrial being) are viewed in popular piety—and somewhat in theological reflection—as malevolent in intent.

Angels are generally grouped in orders of four, six, or seven in the first ranks, of which there may be several. The use of four, which symbolically implies perfection and is related to the four cardinal points, is found in Judaism, Christianity, and Islām. Early Zoroastrianism, much influenced by the astronomical and astrological sciences of ancient Iran, coordinated the concept of the seven known planetary spheres with its belief in the heptad (grouping of seven) of celestial beings—*i.e.*, the *amesha spentas* of Ahura Mazdā: Spenta Mainyu (the Holy Spirit), Vohu Mana (Good Mind), Asha (Truth), Ārmaiti (Right Mindedness), Khshathra (Kingdom), Haurvatāt (Wholeness), and Ameretāt (Immortality). In later Zoroastrianism, though not in the *Gāthās* (the early hymns, believed to have been written by Zoroaster, in the Avesta, the sacred scriptures), Ahura Mazdā and Spenta Mainyu were identified with each other, and the remaining bounteous immortals were grouped in an order of six. Over against the bounteous immortals, who helped to link the spiritual and the material worlds together, was the counterpart of the Holy Spirit, namely Angra Mainyu, the Evil Spirit, who later became the great adversary Ahriman (the prototype of the Jewish, Christian, and Islāmic Satan), and the daevas, who were most likely gods of early Indo-Iranian religion. Allied with Angra Mainyu against Ahura Mazdā were Akōman (Evil Mind), Indrā-vāyū (Death), Saurva (a daeva of death and disease), Nāŋhaithya (a daeva related to the Vedic god Nāsatya), Tauru (difficult to identify), and Zairi (the personification of Haoma, the sacred drink related to the sacrifices of both *ahuras* and daevas). Among other demonic figures is Aēshma (violence, fury, or the aggressive impulse that consumes man)—who may well be the demon Asmodeus of the book of Tobit, Āz (Concupiscence or Lust), Mithrāndruj (He Who Lies to Mithra or False Speech), Jēh (the demon Whore, created later by Ahriman to defile the human race), and many others (see also ZOROASTRIANISM AND PARSIISM).

The orders of ranks of angels and demons in Zoroastrianism

Angelology and demonology in Judaism became more highly developed during and after the period of the Babylonian Exile (6th–5th centuries BC), when contacts were made with Zoroastrianism. In the Old Testament, Yahweh is called the Lord of hosts. These hosts (Sabaoth) are the heavenly army that fights against the forces of evil and performs various missions, such as guarding the entrance to Paradise, punishing evildoers, protecting the faithful, and revealing God's Word to man. Two archangels are mentioned in the canonical Old Testament: Michael, the warrior leader of the heavenly hosts, and Gabriel, the heavenly messenger. Two are mentioned in the apocryphal Old Testament: Raphael, God's healer or helper (in the book of Tobit), and Uriel (Fire of God), the watcher over the world and the lowest part of hell (in II Esdras). Though these are the only four named, seven archangels are noted in Tob. 12:15. Besides the archangels, there were also other orders of angels, the cherubim and seraphim, which have been noted earlier.

The angels and demons of Judaism and Christianity

Under the influence of Zoroastrianism, Satan, the adversary, probably evolved into the archdemon. Other demons included Azazel (the demon of the wilderness, incarnated in the scapegoat), Leviathan and Rahab (demons of chaos), Lilith (a female night demon), and others. To protect themselves from the powers of the demons and unclean spirits, Jews influenced by folk beliefs and customs (as with Christians later) often carried charms,

amulets, and talismans inscribed with efficacious formulas (see also JUDAISM).

Christianity, probably influenced by the angelology of Jewish sects such as the Pharisees and the Essenes as well as of the Hellenistic world, further enhanced and developed theories and beliefs in angels and demons. In the New Testament, celestial beings were grouped into seven ranks: angels, archangels, principalities, powers, virtues, dominions, and thrones. In addition to these were added the Old Testament cherubim and seraphim, which, with the seven other ranks, comprised the nine choirs of angels in later Christian mystical theology. Various other numbers of the orders of angels have been given by early Christian writers: four, in *The Sibylline Oracles* (a supposedly Jewish work that shows much Christian influence); six, in the *Shepherd of Hermas*, a book accepted as canonical in some local early Christian churches; and seven, in the works of Clement of Alexandria and other major theologians. In both folk piety and theology the number has generally been fixed at seven. The angels receiving most attention and veneration in Christianity were the four angels mentioned in the Old Testament and the Apocrypha. Michael became the favourite of many, and in the practice of his cult there was often some confusion with St. George, who was also a warrior figure.

Demonology experienced a renewal in Christianity that probably would have been acceptable in Zoroastrianism. Satan, the archenemy of the Christ; Lucifer, the fallen Light Bearer; and the originally Canaanite Beelzebub, the Lord of Flies (or, perhaps, Beelzebul, the Lord of Dung), mentioned by Jesus, are all devils. The concept and term devil are derived from the Zoroastrian concept of *daevas* and the Greek word *daibolos* ("slanderer" or "accuser"), which is a translation of the Jewish concept of Satan. As a singular demonic force or personification of evil, the devil's chief activity was to tempt man to act in such a way that he would not achieve his supraterrestrial destiny. Because demons were believed to inhabit waterless wastelands, where hungry and tired persons often had visual and auditory hallucinations, early Christian monks went into the deserts to be the vanguard of God's army in joining battle with the tempting devils. They often recorded that the devil came to them in visions as a seductive woman, tempting them to violate their vows to keep themselves sexually pure, both physically and mentally.

Worship of demons in the West

During certain periods in Christian Europe, especially the Middle Ages, worship of demons and the practice of witchcraft brought about the wrath of both church and people on those suspected of practicing diabolical rites, such as the Black Mass. One formula from the Black Mass (the mass said in reverse and with an inverted crucifix on the altar) has survived in popular magic: "hocus-pocus," an abbreviated from of "Hoc est corpus meum" ("This is my body"), the words of institution in the Eucharist, or Holy Communion. Witchcraft and sorcery have been closely associated with demonology in the thought of Christianity, especially in the West.

In the second half of the 20th century, in connection with a renewed interest in the supernatural, there has been evidence of a revival of demon worship and black magic, although this has generally been restricted to small cults that have proved to be quite ephemeral.

Angelology and demonology in Islām are closely related to similar doctrines in Judaism and Christianity. Besides the four throne bearers of Allāh, four other angels are well known: Jibrīl (Gabriel), the angel of revelation; Mikāl (Michael), the angel of nature, providing man with food and knowledge; 'Izrā'īl, the angel of death; and Isrāfīl, the angel who places the soul in the body and sounds the trumpet for the Last Judgment. Demons also contend for control of men's lives, the most prominent being Iblīs (the Devil), who tempts mortal man, or Shayṭan, or Satan (see also ISLĀM).

In the religions of the East. As noted earlier, the function of angels in Eastern religions was carried by *avatāras*, *bodhisattvas*, and other such spiritual beings who were extensions of God or the sacred. Belief in demons was and is very widespread, influencing various rituals and practices to counteract the forces that are hostile to man and nature. In Hinduism, the *asuras* (the Zoroastrian *ahuras*) are the demons who oppose the *devas* (the gods). Both vied for the *homa*, or the *amṛta* (the sacred drink that gives power), but the god Viṣṇu (the preserver), incarnated as a beautiful woman (Mohinī), aided the gods so that they alone would drink the *amṛta*, thus giving them power over the demons. Among the various classes of Hindu *asuras* (demons) are *nāgas* (serpent demons); Ahī (the demon of drought); and Kaṃsa (an archdemon).

Types of demons in Eastern religions

Demons that afflict men include the *rākṣasas*, grotesque and hideous beings of various shapes who haunt cemeteries, impel men to perform foolish acts, and attack *sadhus* (saintly men), and *piśacas*, beings who haunt places where violent deaths have occurred. Buddhists often view their demons as forces that inhibit man from achieving Nirvāṇa (bliss or the extinction of desire). Included among such beings are Māra, an arch tempter, who, with his daughters, Rati (Desire), Rāga (Pleasure), and Tanhā (Restlessness), attempted to dissuade Siddhārtha Gautama, the Buddha, from achieving his Enlightenment. As Mahāyāna (Greater Vehicle) Buddhism spread to Tibet, China, and Japan, many of the demons of the folk religions of these areas were incorporated into Buddhist beliefs. The demons of Chinese religions, the *kuei-shen*, are manifested in all aspects of nature. Besides these nature demons there are goblins, fairies, and ghosts. Because the demons were believed to avoid light, the Chinese who were influenced by Taoism and folk religions used bonfires, firecrackers, and torches to ward off the *kuei*. Japanese religions are similar to Chinese religions in the multiplicity of demons with which men must contend. Among the most fearsome of the Japanese demons are the *oni*, evil spirits with much power, and the *tengu*, spirits that possess man and that generally must be exorcized by priests (see also HINDUISM; BUDDHISM; JAINISM).

In nonliterate religions. The spiritual beings of nonliterate religions of Asia, Africa, Oceania, and the Americas are generally viewed as malevolent or benevolent according to circumstances rather than because of their inherent nature. Eshu, a god of the Yoruba of Nigeria, for example, is looked upon as a protective, benevolent spirit as well as a spirit with an evil power that may be directed toward one's enemies. These beings possess what is called mana (supernatural power), a Melanesian term that can be applied both to spirits and to persons of special status, such as chiefs or shamans. In nonliterate religions, the spirits of nature are generally venerated in return for certain favours or to ward off catastrophes, much in the manner of the religion of ancient Rome. Ancestor gods abound, and thus the ghosts of the dead must be placated, often with the performance of elaborate rites (see also SACRED OFFICES AND ORDERS: *Shamanism*; RELIGIOUS AND SPIRITUAL BELIEF: *Ancestor worship*).

Conclusion. Though traditional beliefs in angels and demons have been questioned among those cultures affected by Western science and technology, reinterpretations of such beliefs, under the influence of psychological studies and the study of myth in the history of religions, have been of significance to theological reflection. By viewing angels and demons functionally, rather than in terms of their natures, modern man may discover that he has a greater kinship than he has generally realized with men of previous or different cultures in his attempt to gain an advantageous rapport with the transcendent, social, and psychological realms that he faces in everyday life. (L.F.)

Salvation

Salvation is the deliverance or redemption of man from such fundamentally negative or disabling conditions as suffering, evil, finitude, and death. In some religious beliefs it also entails the restoration or raising up of the natural world to a higher realm or state. The idea of salvation is a characteristic religious notion related to an issue of profound human concern.

NATURE AND SIGNIFICANCE

It could be argued reasonably that the primary purpose of all religions is to provide salvation for their adherents,

and the existence of many different religions indicates that there is a great variety of opinion about what constitutes salvation and the means of achieving it. That the term salvation can be meaningfully used in connection with so many religions, however, shows that it distinguishes a notion common to men and women of a wide range of cultural traditions.

The fundamental idea contained in the English word salvation, and the Latin *salvatio* and Greek *sōtēria* from which it derives, is that of saving or delivering from some dire situation. The term soteriology denotes beliefs and doctrines concerning salvation in any specific religion, as well as the study of the subject. The idea of saving or delivering from some dire situation logically implies that mankind, as a whole or in part, is in such a situation. This premise, in turn, involves a series of related assumptions about human nature and destiny.

Original and ultimate states of man and the world

Objects and goals. The creation myths of many religions express the beliefs that have been held concerning the original state of mankind in the divine ordering of the universe. Many of these myths envisage a kind of Golden Age at the beginning of the world, when the first human beings lived, serene and happy, untouched by disease, aging, or death and in harmony with a divine Creator. Myths of this kind usually involve the shattering of the ideal state by some mischance, with wickedness, disease, and death entering into the world as the result. The Adam and Eve myth is particularly notable for tracing the origin of death, the pain of childbirth, and the hard toil of agriculture, to man's disobedience of his maker. It expresses the belief that sin is the cause of evil in the world, and implies that salvation must come through man's repentance and God's forgiveness and restoration.

In ancient Iran, a different cosmic situation was contemplated, one in which the world was seen as a battleground of two opposing forces: good and evil, light and darkness, life and death. In this cosmic struggle, mankind was inevitably involved, and the quality of human life was conditioned by this involvement. Zoroaster, the founder of Zoroastrianism, called upon men to align themselves with the good, personified in the god Ahura Mazdā, because their ultimate salvation lay in the triumph of the cosmic principle of good over evil, personified in Ahriman. This salvation involved the restoration of all that had been corrupted or injured by Ahriman at the time of his final defeat and destruction. Thus the Zoroastrian concept of salvation was really a return to a Golden Age of the primordial perfection of all things, including man. Some ancient Christian theologians (*e.g.*, Origen) also conceived of a final "restoration" in which even devils, as well as men, would be saved; this idea, called universalism, was condemned by the church as heresy.

In those religions that regard man as essentially a psychophysical organism (*e.g.*, Judaism, Christianity, Zoroastrianism, Islām), salvation involves the restoration of both the body and soul. Such religions therefore teach doctrines of a resurrection of the dead body and its reunion with the soul, preparatory to ultimate salvation or damnation. In contrast, some religions have taught that the body is a corrupting substance in which the soul is imprisoned (*e.g.*, Orphism, an ancient Greek mystical cult; Hinduism; and Manichaeism, an ancient dualistic religion of Iranian origin). In this dualistic view of human nature, salvation has meant essentially the emancipation of the soul from its physical prison or tomb and its return to its ethereal home. Such religions generally explain the incarceration of the soul in the body in terms that imply the intrinsic evil of physical matter. Where such views of human nature were held, salvation therefore meant the eternal beatitude of the disembodied soul.

Resurrection and immortality

Christian soteriology contains a very complex eschatological program (regarding the final end of man and the world), which includes the fate of both individual persons and the existing cosmic order. The return of Christ will be heralded by the destruction of the heaven and earth and the resurrection of the dead. The Last Judgment, which will then take place, will result in the eternal beatitude of the just, whose souls have been purified in purgatory, and the everlasting damnation of the wicked. The saved, reconstituted by the reunion of soul and body, will forever enjoy the Beatific Vision; the damned, similarly reconstituted, will suffer forever in hell, together with the devil and the fallen angels. Some schemes of eschatological imagery, used by both Christians and Jews, envisage the creation of a new heaven and earth, with a New Jerusalem at its centre.

Means. The hope of salvation has naturally involved ideas about how it might be achieved. These ideas have varied according to the form of salvation envisaged; but the means employed can be divided into three significant categories: (1) the most primitive is based on belief in the efficacy of ritual magic—initiation ceremonies, such as those of the ancient mystery religions, afford notable examples; (2) salvation by self-effort, usually through the acquisition of esoteric knowledge, ascetic discipline, or heroic death, has been variously promised in certain religions—Orphism, Hinduism, Islām, for example; and (3) salvation by divine aid, which has usually entailed the concept of a divine saviour who achieves what man cannot do for himself—as in Christianity, Judaism, Islām.

BASIC CONTEXT

Man's awareness of time and death

The cosmic situation. *Time.* Study of the relevant evidence shows the menace of death as the basic cause of soteriological concern and action. Salvation from disease or misfortune, which also figures in religion, is of a comparatively lesser significance, though it is often expressive of more immediate concerns. But the menace of death is of another order, and it affects man more profoundly because of personal awareness of the temporal categories of past, present, and future. This time-consciousness is possessed by no other species with such insistent clarity. It enables man to draw upon past experience in the present and to plan for future contingencies. This faculty, however, has another effect: it causes man to be aware that he is subject to a process that brings change, aging, decay, and ultimately death to all living things. Man, thus, knows what no other animal apparently knows about itself, namely, that he is mortal. He can project himself mentally into the future and anticipate his own decease. Man's burial customs grimly attest to his preoccupation with death from the very dawn of human culture in the Paleolithic Period. Significantly, the burial of the dead is practiced by no other species.

The menace of death is thus inextricably bound up with man's consciousness of time. In seeking salvation from death, man has been led on to a deeper analysis of his situation, in which he has seen in his subjection to time the true cause of the evil that besets him. The quest for salvation from death, accordingly, becomes transformed into one for deliverance from subjugation to the destructive flux of time. How such deliverance might be effected has been conceived in varying ways, corresponding to the terms in which the temporal process is imagined. The earliest known examples occur in ancient Egyptian religious texts. In the so-called Pyramid Texts (*c.* 2400 BC), the dead pharaoh seeks to fly up to heaven and join the sun-god Re on his unceasing journey across the sky, incorporated, thus, in a mode of existence beyond change and decay. A passage in the later Book of the Dead (1200 BC) represents the deceased, who has been ritually identified with Osiris, declaring that he comprehends the whole range of time in himself, thus asserting his superiority to it.

The recognition that mankind is subject to the inexorable law of decay and death has produced other later attempts to explain its domination by time and to offer release from it. Such attempts are generally based on the idea that the temporal process is cyclical, not linear, in its movement. Into this concept, a belief in metempsychosis (transmigration of souls) can be conveniently fitted. For the idea that souls pass through a series of incarnations becomes more intelligible if the process is seen as being cyclical and in accordance with the pattern of time that apparently governs all the forms of being in this world. The conception has been elaborated in various ways in both Eastern and Western religions. In Hinduism and Buddhism, elegantly imaginative chronological systems have been worked out, comprising *mahāyugas*, or periods of 12,000 years, each

Man's deliverance from time and death

year of which represented 360 human years. In turn, 1,000 *mahāyugas* made up one *kalpa,* or one day in the life of Brahmā, and spanned the duration of a world from its creation to its destruction. After a period of quiescence, the world would be re-created by Brahmā for another *kalpa.* The purpose of this immense chronological scheme was to emphasize how the unenlightened soul was doomed to suffer an infinite series of incarnations, with all of their attendant pain of successive births and deaths. In the Orphic texts of ancient Greece, man's destiny to endure successive incarnations is significantly described as "the sorrowful weary Wheel," from which the Orphic initiate hoped to escape through the secret knowledge imparted to him.

Nature. As an alternative interpretation to this view of man's fatal involvement with time, the tragedy of the human situation has also been explained in terms of the soul's involvement with the physical universe. In some systems of thought (*e.g.,* Hinduism and Buddhism), the two interpretations are synthesized; and in such systems it is taught that, by accepting the physical world as reality, the soul becomes subject to the process of time.

Man's deliverance from the body and the natural world

Concentration on the soul's involvement with matter as being the cause of the misery of human life has generally stemmed from a dualistic view of human nature. The drawing of a sharp distinction between spirit and matter has been invariably motivated by a value judgment: namely, that spirit (or soul) is intrinsically good and of transcendent origin, whereas matter is essentially evil and corrupting. Through his body, man is seen to be part of the world of nature, sharing in its processes of generation, growth, decay, and death. How his soul came to be incarcerated in his corruptible body has been a problem that many myths seek to explain. Such explanations usually involve some idea of the descent of the soul or its divine progenitor from the highest heaven and their fatal infatuation with the physical world. The phenomenon of sexual intercourse has often supplied the imagery used to account for the involvement of the soul in matter and the origin of its corruption. Salvation has thus been conceived in this context as emancipation from both the body and the natural world. In Gnosticism and Hermeticism—esoteric theosophical and mystical movements in the Greco-Roman world—and the teaching of St. Paul deliverance was sought primarily from the planetary powers that were believed to control human destiny in the sublunar world.

Man's responsibility. The idea that man is in some dire situation, from which he seeks to be saved, necessarily involves explaining the cause of his predicament. The explanations provided in the various religions divide into two kinds: those that attribute the cause to some primordial mischance and those that hold man to be himself responsible. Some explanations that make man directly responsible represent him also as the victim of the deceit of a malevolent deity or demon.

Because death has been universally feared but rarely accepted as a natural necessity, the mythologies of many peoples represent the primeval ancestors of mankind as having accidentally lost, in some way, their original immortality. One Sumerian myth, however, accounts for disease and old age as resulting from the sport of the gods when they created mankind. In contrast, the Hebrew story of Adam and Eve finds the origin of death in their act of disobedience in eating of the tree of knowledge of good and evil, forbidden to them by their maker. This causal connection between sin and death was elaborated by St. Paul in his soteriology, outlined in his letter to the Romans, and formed the basis of the Christian doctrine of original sin. According to this doctrine, through seminal identity with Adam, every human being must partake of the guilt of Adam's sin, and even at birth, a child is already deserving of God's wrath for its share in the original sin of mankind and before it acquires the guilt of its own actual sin. Moreover, because each individual inherits the nature of fallen humanity, he has an innate predisposition to sin. This doctrine of man means that no person can, by his volition and effort, save himself but depends absolutely upon the saving grace of Christ.

Moral and intellectual defects

Wherever a dualistic view of human nature has been held, it has been necessary to explain how ethereal souls first became imprisoned in physical bodies. Generally, the cause has been found in the supposition of some primordial ignorance or error rather than in a sinful act of disobedience or revolt—*i.e.,* in an intellectual rather than a moral defect. According to the Hindu philosophical system known as Advaita Vedānta, a primordial ignorance (*avidyā*) originally caused souls to mistake the empirical world for reality and so become incarnated in it. By continuing in this illusion, they are subjected to an unceasing process of death and rebirth (*saṃsāra*) and all of its consequent suffering and degradation. Similarly, in Buddhism, a primordial ignorance (*avijjā*) also started the "chain of causation" (*paṭiccasamuppāda*) that produces the infinite misery of unending rebirth in the empirical world.

METHODS AND TECHNIQUES

Ritual. The means by which salvation might be achieved has been closely related to the manner in which salvation has been conceived and to what has been deemed to be the cause of man's need of it. Thus in ancient Egypt, where salvation was from the physical consequences of death, a technique of ritual embalmment was employed. Ritual magic has also been used in those religions that require their devotees to be initiated by ceremonies of rebirth (*e.g.,* Baptism in water in Christianity, in bull's blood in rites of Cybele) and by symbolic communion with a deity through a ritual meal in the Eleusinian Mysteries, Mithraism, and Christianity (communion).

Knowledge. Religions that trace the ills of man's present condition to some form of primordial error, or ignorance, offer knowledge that will ensure salvation. Such knowledge is of an esoteric kind and is usually presented as divine revelation and imparted secretly to specially prepared candidates. In some instances (*e.g.,* Buddhism and Yoga), the knowledge imparted includes instruction in mystical techniques designed to achieve spiritual deliverance.

Devotion and service. Whenever mankind has been deemed to need divine aid for salvation, there has been an emphasis on a personal relationship with the saviour-god concerned. Such relationship usually connotes faith in and loving devotion and service toward the deity, and such service may involve moral and social obligations. Judaism, Christianity, Islām, and the *bhakti* cults of India afford notable examples. Christianity adds a further requirement in this context: because human nature is basically corrupted by sin, God's prevenient (antecedent, activating) grace is needed before man's will can be disposed even to desire salvation.

VARIETIES OF SALVATION IN WORLD RELIGIONS

Ancient Egypt. The Pyramid Texts of ancient Egypt provide the earliest evidence of man's quest for salvation. They reveal that by about 2400 BC a complex soteriology connected with the divine kingship of the pharaohs had been established in Egypt. This soteriology was gradually developed in concept and ritual practice and was popularized; *i.e.,* the original royal privilege was gradually extended to all of the classes of society, until by about 1400 BC it had become an elaborate mortuary cult through which all who could afford its cost could hope to partake of the salvation it offered. This salvation concerned three aspects of postmortem existence, as imagined by the ancient Egyptians, and, in the concept of Osiris, it involved the earliest instance of a saviour-god. An elaborate ritual of embalmment was designed to save the corpse from decomposition and restore its faculties so that it could live in a well-equipped tomb. This ritual imitated the acts that were believed to have been performed by the gods to preserve the body of Osiris, with whom the deceased was ritually assimilated. The next concern was to resurrect the embalmed body of the dead person, as Osiris had been resurrected to a new life after death. Having thus been saved from the consequences of death, the revivified dead had to undergo a judgment (presided over by Osiris) on the moral quality of his life on earth. In this ordeal, the deceased could be saved from an awful second death only by personal integrity. If he safely passed the test, he was declared *maa kheru* ("true of voice") and was admitted to the beatitude of the realm over which Osiris reigned.

Postmortem salvation through Osiris

This Osirian mortuary cult, with its promise of post-mortem salvation, was practiced from about 2400 BC until its suppression in the Christian Era. In some respects, it constitutes a prototype of Christianity as a salvation religion.

Hinduism. Running through the great complex of beliefs and ritual practices that constitute Hinduism is the conviction that the soul or self (*ātman*) is subject to *saṃsāra—i.e.,* the transmigration through many forms of incarnation. Held together with this belief is another, *karman—i.e.,* that the soul carries with it the burden of its past actions—which conditions the forms of its future incarnations. As long as the soul mistakes this phenomenal world for reality and clings to existence in it, it is doomed to suffer endless births and deaths. The various Indian cults and philosophical systems offer ways in which to attain *mokṣa* or *mukti* ("release"; "liberation") from the misery of subjection to the inexorable process of cosmic time. Basically, this liberation consists in the soul's effective apprehension of its essential unity with Brahman, the supreme *Ātman* or essence of reality, and its merging with it. Most of the ways by which this goal may be attained require self-effort in mastering meditation techniques and living an ascetic life. But, in the devotional (*bhakti*) cults associated with Viṣṇu (Vishnu) and Śiva (Shiva), an intense personal devotion to the deity concerned is believed to earn divine aid to salvation.

Buddhism. Buddhism accepts the principles of *saṃsāra* and *karman* (Pāli: *kamma*), but it differs in one important respect from the Hindu conception of man. Instead of believing that an *ātman,* or soul, passes through endless series of incarnations, Buddhism teaches that there is no such preexistent immortal soul that migrates from body to body. Each individual consists of a number of physical and psychic elements (*khandhas*) that combine to create the sense of personal individuality. But this combination is only temporary and is irreparably shattered by death, leaving no element that can be identified as the soul or self. By a subtle metaphysical argument, however, it is maintained that the craving for personal existence generated by the *khandhas* causes the birth of another such personalized combination, which inherits the *karma* of a sequence of previous combinations of *khandhas.*

Salvation through Enlightenment or *bodhisattvas*

The Enlightenment won by Gautama Buddha was essentially about the cause of existence in the phenomenal world, from which suffering inevitably stemmed. Buddhist teaching and practice have, accordingly, been designed to acquaint men with their true nature and situation and enable them to free themselves from craving for existence in the space-time world and so achieve Nirvāṇa. Traditionally, this goal has been presented in negative terms—as the extinction of desire, attachment, ignorance, or suffering—creating the impression that Buddhist salvation means the complete obliteration of individual consciousness. In one sense, this is so; but, in terms of Buddhist metaphysics, ultimate reality transcends all the terms of reference relevant to existence in this world.

Theoretically, the Buddhist initiate should, by his own effort in seeking to eradicate desire for continued existence in the empirical world, achieve his own salvation. But, as Buddhism developed into a popular religion in its Mahāyāna ("Greater Vehicle") form, provision was made for the natural human desire for assurance of divine aid. Consequently, belief in many saviours, known as *bodhisattvas* ("Buddhas-to-be"), developed, together with elaborate eschatologies concerning human destiny. According to these, before the ultimate achievement of Nirvāṇa, the faithful could expect to pass through series of heavens or hells, according to their merits or demerits and the intensity of their devotion to a *bodhisattva.*

Judaism. Because Judaism is by origin and nature an ethnic religion, salvation has been primarily conceived in terms of the destiny of Israel as the elect people of Yahweh, the God of Israel. It was not until the 2nd century BC that there arose a belief in an afterlife, for which the dead would be resurrected and undergo divine judgment. Before that time, the individual had to be content that his posterity continued within the holy nation. But, even after the emergence of belief in the resurrection of the dead, the essentially ethnic character of Judaism still decisively influenced soteriological thinking. The apocalyptic faith, which became so fervent as Israel moved toward its fateful overthrow by the Romans in AD 70, conceived of salvation as the miraculous intervention of Yahweh or his Messiah in world affairs. This saving act would culminate in the Last Judgment delivered on the nations that oppressed Israel and Israel's glorious vindication as the people of God. From the end of the national state in the Holy Land in AD 70, Jewish religion, despite the increasing recognition of personal significance, has remained characterized by its essential ethnic concern. Thus, the Exodus from Egypt has ever provided the typal imagery in terms of which divine salvation has been conceived, its memory being impressively perpetuated each year by the ritual of the Passover. The restoration of the holy nation, moreover, always has been linked with its Holy Land; and Hebrew literature, both in biblical and later forms, has lovingly described the establishment of a New Jerusalem and a new Temple of Yahweh ("the Lord"), whether it be in this world or in some new cosmic order. Into this new order, the rest of mankind, repentant and purified, will be incorporated; for the original promise made to the patriarch Abraham included all men within the divine blessing. In the Book of Zechariah, the ultimate salvation of mankind is graphically envisaged: the Gentiles, in company with the Jews, will return to serve God in an ideal Jerusalem.

Salvation through Yahweh's action and power

Christianity. Christianity has been described as the salvation religion par excellence. Its primary premise is that the incarnation and sacrificial death of its founder, Jesus Christ, formed the climax of a divine plan for mankind's salvation. This plan was conceived by God consequent on the Fall of Adam, the progenitor of the human race, and it would be completed at the Last Judgment, when the Second Coming of Christ would mark the catastrophic end of the world. This soteriological evaluation of history finds expression in the Christian division of time into two periods: before Christ (BC) and Anno Domini (AD)—*i.e.,* the years of the Lord.

Salvation through Christ

The evolution of the Christian doctrine of salvation was a complicated process essentially linked with the gradual definition of belief in the divinity of Jesus of Nazareth. In Christian theology, therefore, soteriology is an integral part of what is termed Christology. Whereas the divinity of Jesus Christ has been the subject of careful metaphysical definition in the creeds, the exact nature and mode of salvation through Christ has not been so precisely defined. The church has been content to state, in its creeds, that Christ was incarnated, crucified, died, and rose again "for us men, and for our salvation."

The basic tenets of Christian soteriology may be summarized as follows: man is deserving of damnation by God for the original sin, which he inherits by descent from Adam, and for his own actual sin. But, because sin is regarded as also putting man in the power of the devil, Christ's work of salvation has been interpreted along two different lines. Thus, his crucifixion may be evaluated as a vicarious sacrifice offered to God as propitiation or atonement for human sin. Alternatively, it may be seen as the price paid to redeem man from the devil. These two ways of interpreting the death of Christ have provided the major themes of soteriological theory and speculation in Christian theology. Despite this fluidity of interpretation, belief in the saving power of Christ is fundamental to Christianity and finds expression in every aspect of its faith and practice.

Islām. Muḥammad regarded himself as "a warner clear" and as the last and greatest of a line of prophets whom Allāh had sent to warn his people of impending doom. Although the word *najāt* (Arabic: "salvation") is used only once in the Qur'ān, the basic aim of Islām is salvation in the sense of escaping future punishment, which will be pronounced on sinners at the Last Judgment. Muḥammad did teach that Allāh had predestined some men to heaven and others to hell; but the whole logic of his message is that submission to Allāh is the means to salvation, for Allāh is merciful. Indeed, faithful submission is the quintessence of Islām, the word Islām itself meaning submission. Although in his own estimation Muḥammad was the

prophet of Allāh, in later Muslim devotion he came to be venerated as the mediator between God and man, whose intercession was decisive.

Zoroastrianism and Parsiism. According to Zoroaster, a good and evil force struggled for mastery in the universe. Man had to decide on which side to align himself in this fateful contest. This dualism was greatly elaborated in later Zoroastrianism and Parsiism, which derived from it. Good, personified as the god Ormazd, and evil, as the demonic Ahriman, would contend for 12,000 years with varying fortune. At last Ormazd would triumph, and Saoshyans, his agent, would resurrect the dead for judgment. The righteous would pass to their reward in heaven, and the wicked be cast into hell. But this situation was of temporary duration. A meteor would later strike the earth, causing a flood of molten metal. Through this flood all would have to pass as an ordeal of purgation. The sensitivity of each to the anguish would be determined by the degree of his guilt. After the ordeal, all men would become immortal, and all that Ahriman had harmed or corrupted would be renewed. Salvation thus took the form of deliverance from postmortem suffering; for ultimate restoration was assured to all after suffering the degree of purgation that the nature of their earthly lives entailed. (S.G.F.B.)

Providence

Providence is the quality in divinity on which man bases his belief in a benevolent divine intervention in human affairs and the affairs of the world he inhabits. The forms that this belief takes differ, depending on the context of the religion and the culture in which they function.

In one view the concept of Providence, divine care of man and the universe, can be called the religious answer to man's need to know that he matters, that he is cared for, or even that he is threatened, for in this view all religions are centred on man, and man is individually and collectively in constant need of reassurance that he is not an unimportant item in an indifferent world; if he cannot be comforted, to be threatened is better than to be alone in an empty void of nothingness. According to J. van Baal, a Dutch anthropologist,

> Man experiences his universe as a universe full of intentions, a universe which holds a claim on him, addressing him with something undefined, urging him to act or to be in some way or another. The experience is strongest in moments of crisis, when events turn up with such an overwhelming force that it is as if they address their victim, delivering a message to him.

In answer to such a universe, religions must offer a coherent view of God or gods, world, and mankind and must give man and his physical or psychical well-being, or both, a prominent place within this world view. Thus, in all religions Divine Providence or its equivalent is an element of some importance.

NATURE AND SIGNIFICANCE

Basic forms of Providence. Basically, there are two possible forms of belief in Providence. In the first, man believes in more or less divine beings that are responsible for the world generally and for the welfare of man specifically. Although omnipotence as an attribute of gods is rare, it is true that, as a rule, gods and other divine beings have considerable power not only over man but also over nature. The gods take care of the world and of mankind, and their intentions toward mankind are normally positive. The capricious and arbitrary gods of paganism exist for the most part only in the imagination of those Christian theologians who attempt to denigrate the pagan religions. Gods and men are generally connected into one community by reciprocal duties and privileges. The belief in evil spirits does not contradict this belief in Providence but, on the contrary, strengthens it, just as in Christianity the belief in the devil might serve to strengthen the belief in God.

Belief in a cosmic order

In the second form, man believes in a cosmic order in which the welfare of man has its appointed place. This cosmic order is usually conceived as a divine order that is well intentioned toward man and is working for man's well-being as long as he is willing to insert himself into this order, to follow it willingly, and not to upset it by perversion or rebellion; the firmness of the order, however, may become inexorable and thus lead to fatalism, the belief in an impersonal destiny against which man is powerless. In that case a clash between the concepts of Providence and fatalism is inevitable. In most religions, however, both views are combined in some way.

Etymological history of the term Providence. The English word Providence is derived from the Latin term *providentia,* which primarily means foresight or foreknowledge but also forethought and Providence in the religious sense; thus, Cicero used the phrase the "Providence of the gods" (*deorum providentia*). The Stoic philosophers thoroughly discussed the significance of the term Providence, and some of them wrote treatises on the subject. A hymn to Zeus written about 300 BC by Cleanthes, a Greek poet and philosopher, is a glorification of the god as a benevolent and foreseeing ruler of the world and of mankind. According to Cleanthes, God has planned the world in accordance with this Providence:

> For thee this whole vast cosmos, wheeling round
> The earth, obeys, and where thou leadest
> It follows, ruled willingly by thee.

The author asserts that "naught upon Earth is wrought in thy despite, O God" and that in Zeus all things are harmonized. Seneca, a Roman Stoic philosopher, formulates the belief in Providence in one of his dialogues as follows: man should believe "that Providence rules the world and that God cares for us." The Stoic school disagreed with those who believed that the world was ruled by blind fate; they did not deny that a controlling power exists, but, as everything happens according to a benevolent divine plan, they preferred to call this power Providence. According to the Stoic emperor Marcus Aurelius, God wills everything that happens to man, and for that reason nothing that occurs can be considered evil. Stoic ideas about Providence influenced Christianity.

In later Latin after the emperor Augustus, the word Providence was used as a designation of the deity. Seneca, for example, wrote that it is proper to apply the term Providence to God. Finally, Providence was personified as a proper goddess in her own right by Macrobius, a Neoplatonic Roman author, who wrote in defense of paganism about 400.

Epicurus, a 4th–3rd-century-BC Greek philosopher, contested the Stoic belief in Divine Providence, but the objections of his followers could not change the spiritual climate of the Greco-Roman world. More eloquent, perhaps, than the dissertations of the learned Stoic philosophers were the many stories found in a work by Aelian, an early 3rd-century-AD Roman rhetorician, about strange events and miraculous occurrences ascribed to Providence. Aelian, however, was more interested in sensational stories than in historic accuracy.

Latin meanings of the word

The several meanings of the Latin word *providentia* exactly mirror those of its Greek equivalent, *pronoia.* Herodotus, the historian of the 5th century BC, was the first Greek author to use the word in a religious sense when he mentioned Divine Providence as the source of the wisdom that keeps nature in balance and prevents one kind of creature from prevailing over all others. Writers such as the historian Xenophon and the biographer Plutarch used the word for the watchful care of the gods over mankind and the world.

The belief in the existence of a blind and inexorable fate can lead to a conflict with the belief in a benevolent Providence. In the Greco-Roman world, where fatalistic belief was strong and where it found a popular expression in astrology, the belief that the whole world, but particularly man, is governed by the stars was contested by Judaism and Christianity. The Talmud, the authoritative collection of Jewish tradition, teaches that Israel is subject to no star but only to God. An example of this conflict is also found in the novel *The Golden Ass* by Apuleius, a 2nd-century-AD philosopher and rhetorician deeply interested in Hellenistic mystery cults, which taught a faith that liberated man from the power of the stars. In the novel the hero is converted to the goddess Isis; then, the priest of the goddess addresses him:

"Lucius, my friend," he said, "you have endured and performed many labours and withstood the buffetings of all the winds of ill luck. Now at last you have put into the harbour of peace and stand before the altar of loving-kindness. Neither your noble blood and rank nor your education sufficed to keep you from falling a slave to pleasure; youthful follies ran away with you. Your luckless curiosity earned you a sinister punishment. But blind Fortune, after tossing you maliciously about from peril to peril has somehow, without thinking what she was doing, landed you here in religious felicity. Let her begone now and fume furiously wherever she pleases, let her find some other plaything for her cruel hands. She has no power to hurt those who devote their lives to the honour and service of our Goddess's majesty."

Christian use of the term

The Christian use of the term Providence, besides being profoundly influenced by Greek and Roman thought, is based on the Old Testament story of the patriarch Abraham's sacrifice of his son Isaac, which is found in the book of Genesis. Abraham tells Isaac, "God will provide himself with a young beast for a sacrifice, my son." The Hebrew language lacks a proper word to express the notion of Providence, but the concept is well known in the Old Testament.

In the New Testament the word *pronoia* and related words are used rarely, but in no case are they used in the later Christian sense of Providence. This is of interest because the idea of Providence as such is far from foreign to the religious thinking of the New Testament. In the Gospel According to Matthew, for example, Jesus says:

> Are not two sparrows sold for a penny? And not one of them will fall to the ground without your Father's will. But even the hairs of your head are numbered. Fear not, therefore; you are of more value than many sparrows.

Providence as used in Christianity is thus a dogmatic term rather than a biblical term; it indicates that God not only created the world but also governs it and cares for its welfare. A well-known German reference work, *Religion in Geschichte und Gegenwart* ("Religion in History and the Present"), gives a more elaborate and more theological definition of Providence:

> God keeps the world in existence by his care, he rules and leads the world and mankind deliberately according to his purpose, and he does this in his omnipotence as God the Creator, in his goodness and love as revealed by his son Jesus Christ, and to further the salvation of mankind through the Holy Spirit.

BASIC CONCEPTS AND SCOPE

Qualities of the divinity. The concept of Providence is rooted in the belief in the existence of a benevolent, wise, and powerful deity or a number of beings that are benevolent and that are either fully divine or, at least, appreciably wiser and more powerful than man (*e.g.,* ancestors in many religions). Benevolence is the primary requirement. In northern Malawi, death in later life is usually ascribed to the will of the ancestors, but a miscarriage or the death of a very young child is not considered to be their work because such an act would be in contradiction with their benevolent and helpful attitude toward their offspring. The three attributes, however, are all essential for the concept of Providence: the divine being or beings must be well intentioned toward man, must have the necessary wisdom to know what is good for mankind, and must have the power to act on this intention and insight. Benevolence does not exclude the possibility of punishment in cases of transgression. There is probably no god in existence who only rewards and helps and never punishes his believers.

Providence, however, need not operate in a direct way; it may operate through many intermediary beings—*e.g.,* the ancestors and various kinds of spirits in several nonliterate religions or the angels in Christian and Muslim belief—or the concept may be implicit in and expressed by a fixed world order, a cosmic order that makes human life possible biologically, socially, and spiritually and that guarantees its existence in the future. Thus, Providence may become a more or less impersonal principle of cosmic order as instituted and maintained by a divine being, but, if the starting point of a benevolent and just divine being is completely lost sight of or if it is consciously denied, then Providence becomes fate.

Cosmic order. *Notion of cosmic order.* Although the introduction of intermediary beings brings no essential change in the idea of Providence as the divine watchful care for the benefit of mankind, the notion of a cosmic order changes the picture profoundly. Even if the cosmic order is conceived as a benevolent order in which man is able to feel safe and whose very existence reassures him, such an order is different from the personal relationship between man and his god or gods. The concept of an unchangeable world order requires a different reaction. A personal god may, perhaps, be moved by prayer and sacrifice to give or to prevent events; when the order of the world is fixed, however, the course of events cannot be changed by these or any other means. There is probably no religion that acknowledges an all-embracing world order without any exceptions at all. Generally, human beings have such an important function in the order of the world that they also have a certain opportunity to manipulate this order, at least to a certain extent, for instance, by sacrifice or other ritual acts. One opening is presented by the fact that the cosmic order is valid for everything of a more general character, but as a rule the divine will or the free will of man or chance operates on the level of the common occurrences and daily life of the individual. Though in theory the order may govern everything, a large field is left open for different concepts to function. In some cases even uncertainty and chance have their proper place within a determined order. In Yoruba religion (Nigeria), for example, the god Eshu represents the principle of chance and uncertainty and of all that cannot be foreseen. He is one of the gods of the pantheon and has his own sanctuaries and priests.

Manipulation of order and predestination

Another possibility for combining the idea of a personal divine will with a fixed course of events is the concept of predestination best known from Islām and some forms of Calvinism (derived from the thought of John Calvin, a 16th-century French Protestant Reformer) and also important in the theology of Augustine of Hippo, a 4th–5th-century Church Father. Although predestination essentially is concerned with salvation—the question of whether a certain individual will be saved or damned—it is a concept that easily lends itself to a more general application. In a few religions the idea that the individual chooses his own destiny before birth is encountered; *e.g.,* the Batak of Sumatra and some West African tribes. In this conception free will and predestination merge.

In all religions that acknowledge the existence of a more or less impersonal cosmic order, man is expected to work with the cosmos, to insert himself into the cosmic order. Man's behaviour in all fields is governed by a set of rules that are all based on the same principle: to act and to be in harmony with the order of the world, which is natural and divine at the same time.

The cosmic order is given with the creation of the world, but it is possible to question the relation of the Creator to the world after creation. On one hand, there is the belief that God will not abandon the world he has created; on the other, the belief that God created the world and the cosmic order in such a manner that to a great extent the course of the world is fixed from the first beginning and he is no longer involved in it. The latter was, in fact, the thesis of the 17th- and 18th-century Deists in Europe (see RELIGIOUS AND SPIRITUAL BELIEF: *Deism*). The fact of creation helps man to believe in Providence because it would be inconsistent for the creator god or gods not to care for the further existence of the created world. Only persistent disobedience and open rebellion can then furnish a reason for the Creator to abandon or destroy the world. This situation is expressed in the myths of a great flood or some other form of destruction sent as a punishment. There is, however, never a total destruction of the world in these myths, although this final solution may be threatened for the eschatological (ultimate end) future. It may also be promised, if the eschatological events are construed as the definitive institution of a world order that is perfect for all eternity and will never deteriorate.

Cosmic order and ethical principles

The cosmic order is often clearly contrasted with the disorder of chaos. The cosmic order is a total order; it comprises not only all natural things but also social and

ethical rules. This does not mean that cultures and religions centred on a cosmic order have no clear idea of distinctive ethical principles but that ethics is considered as one function of the total cosmic order and as such can never be quite independent. The rules of ethics depend on and are derived from the more general rules that govern the cosmos in its totality; they are no more than special manifestations of these general rules. An example of this attitude can be found in the Greek hymns in praise of the goddess Isis. She is honoured as the queen of heavens; she divided the earth from the heaven, showed the stars their paths, and ordered the course of the sun and the moon. But the same hymn says that she ordained that children should love their parents, that she taught men to honour the images of the gods, and that she made justice stronger than gold and silver. She established penalties for the people practicing injustice and taught that men should have mercy with suppliants. She is also praised because she invented writing, devised marriage contracts, invented navigation, and watches over all men who sail on the sea.

Personal and impersonal forms. The cosmic order can appear in a personalized form, as, for example, the Egyptian goddess Maat; but this personification of the cosmic order is not general: the Iranian Asha, the Indian *ṛta*, and the Chinese Tao are all to a high degree impersonal. Maat represents truth and order; her domain includes not only the order of the nature, but also the social and ethical orders. She plays an important role in the judgment of the dead: the heart of the deceased is weighed against the truth of Maat. She is often called the daughter of Re. In this case, Re is the creator god who not only created the world but also founded the cosmic order as represented by Maat. Her importance is also apparent in the conception of the Maat sacrifice. In Egypt sacrifice is not so much a gift of men to the gods as a sacral technique that enables man to contribute to the maintenance and, if necessary, the restoration of harmony and order in the world. Not only must man live according to Maat but also the gods must live by her truth and order; according to Egyptian texts, the goddess Maat is the food by which the gods live.

Asha, *ṛta*, and Tao

The idea of a determined cosmic order that is natural as well as ethical is an important concept in the Persian religion of Zoroastrianism (also called Mazdaism and, in India, Parsiism) founded during the late 7th and early 6th centuries BC by Zoroaster (Zarathustra). This idea is called Asha and is the counterpart of Drug, which represents evil and deceit and the disorder connected with these. Asha is connected with the sacred element fire. The Indian concept of *ṛta* forms the Indian counterpart of Asha. The gods, especially the Ādityas, protect the world against chaos and ignorance and maintain the world order, which, however, exists independently from the gods. Although the power of *ṛta* operates according to its own principles and laws, man is able, provided he knows the right methods, to manipulate this power to some extent for his own benefit. The proper means for this manipulation is found especially in older Hindu sacrifice. The gods are generally benevolent and friendly toward men who follow *ṛta*, and they punish their own enemies and those of the world order, which in India, too, embraces the social ethical rules.

The concept of Tao is of great importance in Chinese religion, especially in Taoism, founded by Lao-tzu according to tradition in the 6th century BC. Lao-tzu is the author of the *Tao-te Ching* ("Classic of the Way and Its Power") in which he expounds this concept in a manner that is more mystical than philosophical. Tao, literally translated "road," is a difficult and complex concept. It certainly represents the cosmic order, but in Taoism it is even more than that. It is also the concept that gives existence meaning; it is the primeval power that forms the foundation of all that is; and, in some cases, it is even used to designate some kind of high god. Taoism is a mystic religion, and the *Tao-te Ching* is a mystic treatise in which the essence of the Tao is expounded in many parables and metaphors because it cannot be expressed rationally.

Many related concepts exist. The Greek Moira, for instance, is comparable to Asha and *ṛta:* it lacks, however, the mystic overtones of Tao. The Moira in classical Greek religion is not yet fate as this idea was found in Greco-Roman times. The concept of cosmic order may function either in a religious or in a philosophic context; *e.g.*, the pre-established harmony (*harmonia praestabilita*) in the philosophy of Gottfried Wilhelm Leibniz, a German Rationalist, is the cosmic order that holds together and unifies the innumerable individual units, called monads by Leibniz.

Particular objects of Providence. Although cosmic order is necessarily a general idea comprising the whole of the world and all that exists in it, the concept of Providence may be more particular: the benevolent aspect of Providence may be confined to a special group of people or at least be specially related to that group; or a number of patron gods or saints may watch over some specific activity or smaller group. This accounts for the idea of a chosen people watched over and led by a just and loving God. The ancient people of Israel is, perhaps, the best known example; the concept, however, is widespread. Patron gods and patron saints who are particularly charged with caring for some small group, craft, or activity or who operate in special circumstances, such as during illness or war, occur in most religions and are popular in many.

Although Providence in most religions operates primarily for the welfare and the salvation of the community as a whole, it may also be experienced as personal guidance. This latter phenomenon is common in some diverse cultures—*e.g.*, that of the Plains Indians of North America and in some forms of Protestantism in which generally each person is expected to have a private experience of divine guidance. In other cultures and religions, personal guidance is often a prerogative of some person or persons singled out for some reason by God or the gods.

CRITICAL PROBLEMS

It is clear that the concept of Providence by its central position in many religions is connected with numerous other aspects of religion. In monotheistic religions Providence is a quality of the one divinity; in polytheistic religions it may be either a quality of one or more gods or it may be conceived as an impersonal world order on which the gods, too, more or less depend. In the latter case, Providence may lose its aspect of benevolence and become inexorable fate or fickle chance. Most religions show a certain ambivalence; for fate and Providence do not always form a clear-cut contradiction.

Still another form of ambivalence occurs between fate or divine will and the will of man when the latter is conceived as free, or at least free to a certain degree. In some religions the benevolent aspect of Providence appears as grace, and a discussion may arise about the relationship between free will and grace. Perhaps the most difficult problem connected with the notion of Providence is the existence of evil; men have perennially coped with the question of how to reconcile the idea of a provident God or gods with the evident existence of evil in the world. (T.P.v.B.)

The problem of evil

Revelation

Revelation is a religious term that designates the disclosure of divine or sacred reality or purpose to men. In the religious view, such disclosure may come through mystical insights, historical events, or spiritual experiences that transform the lives of individuals and groups.

NATURE AND SIGNIFICANCE

Every great religion acknowledges revelation in the wide sense that its followers are dependent on the privileged insights of its founder or of the original group or individuals with which the faith began. These profound insights into the ultimate meaning of life and the universe, which have been handed down in religious traditions, are arrived at, it is believed, not so much through logical inference as through sudden, unexpected illuminations that invade and transform the human spirit. Those religions that look upon God as a free and personal spirit distinct from the world accept revelation in the more specific sense of a divine self-disclosure, which is commonly depicted on the model of human intersubjective relationships. In the "prophetic" religions (Judaism, Christianity, Islām, and Zoroastrian-

ism), revelation is conceived as a message communicated by God to an accredited spokesman, who is charged to herald the content of that message to an entire people. Revelations received on behalf of the whole community of the faithful are often called "public" (as opposed to "private" revelations, which are given for the guidance or edification of the recipient himself).

The media of revelation

The media by which revelation occurs are variously conceived. Most religions refer to signs, such as auditory phenomena, subjective visions, dreams, and ecstasies. In primitive religions, revelation is often associated with magical techniques of divination. In the prophetic religions, revelation is primarily understood as the "Word of God," enabling the prophet to speak with certainty about God's actions and intentions. In mystical religion (*e.g.*, Islāmic Ṣūfism, Tantric Buddhism) revelation is viewed as an ineffable experience of the transcendent or the divine.

TYPES AND VARIATIONS

Religions of nonliterate cultures. In nonliterate culture revelation is frequently identified with the experience of supernatural power (mana) in connection with particular physical objects, such as stones, amulets, bones of the dead, unusual animals, and other objects. The sacred or holy is likewise believed to be present in sacred trees, groves, shrines, and the like and in elemental realities such as earth, water, sky, and the heavenly bodies. Once specified as holy, such objects take on symbolic value and become capable of mediating numinous (spiritual) experiences to the adherents of a cult. Certain charismatic individuals, such as shamans, who are believed to be in communion with the sacred or holy, perform functions akin to those of the prophet and the mystic in more developed religions.

Religions of the East. Eastern religions are concerned with man's struggle to understand and cope with the predicament of his existence in the world and to achieve emancipation, enlightenment, and unity with the Absolute. Western religions, on the other hand, lay more stress on man's obedient response to the sovereign Word of God. The notion of revelation in the specific sense of a divine self-communication is more apparent in Western than in Eastern religions.

Hinduism. In Hinduism, the dominant religion of India, revelation is generally viewed as a process whereby the religious seeker, actuating his deeper spiritual powers, escapes from the world of change and illusion and comes into contact with ultimate reality. The sacred books are held to embody revelation insofar as they reflect the eternal and necessary order of things.

A major form of Hindu thought, Vedānta, includes two main tendencies: the monistic (*advaita*) and the theistic (*bhakti*). The leading sage of Advaita Vedānta, Śaṅkara (early 9th century), while acknowledging in principle the possibility of coming to a knowledge of the Supreme Reality (Brahman) through inner experience and contemplation of the grades of being, held that in practice a vivid apprehension of the divine arises from meditation on the sacred books, especially the *Upaniṣads*. In Bhakti, systematized by the philosopher Rāmānuja (*c.* 1050–1137), the Absolute is regarded as personal and compassionate. Revelation, consequently, is viewed as the gracious self-manifestation of the divine to those who open themselves in loving contemplation. The devotional theism of Bhakti, very influential in modern India, resembles the pietism and mysticism of the Western religions.

Buddhism. Buddhism, the other great religion originating on Indian soil, conceives of revelation not as a personal intervention of the Absolute into the worldly realm of relativities but as an enlightenment gained through discipline and meditation. Gautama the Buddha (6th to 5th century BC), after a striking experience of human transitoriness and a period of ascetical contemplation, received an illumination that enabled him to become the supreme teacher for all his followers. Although Buddhists do not speak of supernatural revelation, they regard the Buddha as a uniquely eminent discoverer of liberating truth. Some venerate him, some worship him, and all Buddhists seek to imitate him as the most perfect embodiment of ideal manhood—an ideal that he in some way "reveals."

Role of the Buddha

Chinese religions. Chinese wisdom, more world-affirming than the ascetical religions of India, accords little or no place to revelation as this term is understood in the Western religions, though Chinese traditions do speak of the necessity of following a natural harmony in the universe. Taoism, perhaps the most characteristic Chinese form of practical mysticism, finds revelation only in the transparency of the immanent divine principle or way (Tao). Confucianism, while not incompatible with Taoism, is oriented less toward natural mysticism and more toward social ethics and decorum, though it too is concerned with accommodating life to a balance in the natural flow of existence. Confucius (551–479 BC), who refined the best moral teachings that had come down in the tradition, was neither a prophet appealing to divine revelation nor a philosopher seeking to give reasons for his doctrine.

Religions of the West. In the three great religions of the West—Judaism, Christianity, and Islām—revelation is the basic category of religious knowledge. Man knows God and his will because God has freely revealed himself—his qualities, purpose, or instructions.

Judaism. The Israelite faith looked back to the Pentateuch (the first five books of the Old Testament) for its fundamental revelation of God. God was believed to have revealed himself to the patriarchs and prophets by various means not unlike those known to the primitive religions —theophanies (visible manifestations of the divine), dreams, visions, auditions, and ecstasies—and also, more significantly, by his mighty deeds, such as his bringing the Israelites out of Egypt and enabling them to conquer the Holy Land. Moses and the prophets were viewed as the chosen spokesmen who interpreted God's will and purposes to the nation. Their inspired words were to be accepted in loving obedience as the Word of God.

Revelation through events

Rabbinic Judaism, which probably originated during the Babylonian Exile and became organized after the destruction of the Temple by the Romans, concerned itself primarily with the solution of legal and ethical problems. It gradually developed an elaborate system of casuistry resting upon the Torah (the Law, or the Pentateuch) and its approved commentaries, especially the Talmud (commentaries on the Torah), which was regarded by many as equal to the Bible in authority. Orthodox Judaism still recognizes these authoritative sources and insists on the verbal inspiration of the Bible, or at least of the Pentateuch.

Christianity. The New Testament took its basic notions of revelation from the contemporary forms of Judaism (1st century BC and 1st century AD)—*i.e.*, from both normative rabbinic Judaism and the esoteric doctrines current in Jewish apocalyptic circles in the Hellenistic world. Accepting the Hebrew Scriptures as preparatory revelation, Christianity maintains that revelation is brought to its unsurpassable climax in the person of Jesus Christ, who is God's own Son (Heb. 1:1–2), his eternal Word (John 1:1), and the perfect image of the Father (Col. 1:15). The Christian revelation is viewed as occurring primarily in the life, teaching, death, and Resurrection of Jesus, all interpreted by the apostolic witnesses under the illumination of the Holy Spirit. Commissioned by Jesus and empowered by the divine spirit, the apostles, as the primary heralds, hold a position in Christianity analogous to that of the prophets in ancient Israel.

The Apostle Paul, though not personally a witness to the public life of Jesus, is ranked with the Apostles by reason of his special vision of the risen Christ and of his special call to carry the Gospel to the Gentiles. In his letters, Paul emphasized the indispensability of missionary preaching in order that God's revelation in Christ be communicated to all the nations of the world (Rom. 10:11–21).

Christianity has traditionally viewed God's revelation as being complete in Jesus Christ, or at least in the lifetime of the Apostles. Further development is understood to be a deeper penetration of what was already revealed, in some sense, in the 1st century. Periodically, in the course of Christian history, there have been sectarian movements that have attributed binding force to new revelations occurring in the community, such as the 2nd-century Montanists (a heretical group that believed they were of the Age of the Holy Spirit), the 13th century Joachimites (a

mystical group that held a similar view), the 16th-century Anabaptists (radical Protestant sects), and the 17th-century Quakers. In the 19th century the Church of Jesus Christ of Latter-day Saints (popularly known as Mormons) recognized, alongside the Bible, additional canonical scriptures (notably, the Book of Mormon) containing revelations made to the founder, Joseph Smith.

Islām. Islām, the third great prophetic religion of the West, has its basis in revelations received by Muḥammad (*c.* 7th century AD). These were collected shortly after his death into the Qur'ān (Koran), which is regarded by Muslims as the final, perfect revelation—a human copy of the eternal book, dictated to the Prophet. While Islām accords prophetic status to Moses and Jesus, it looks upon the Qur'ān as a correction and completion of all that went before. More than either Judaism or Christianity, Islām is a religion of the Book. Revelation is understood to be a declaration of God's will rather than his personal self-disclosure. Insisting as it does on the absolute sovereignty of God, on man's passivity in relation to the divine, and on the infinite distance between creator and creature, Islām has sometimes been inhospitable to philosophical speculation and mystical experience. Yet in medieval Islām there was both a remarkable flowering of Arabic philosophy and the intense piety of the mystical Ṣūfīs. The rationalism of some philosophers and the theosophical tendencies of some of the Ṣūfīs came into conflict with official orthodoxy.

Revelation through a sacred book

Zoroastrianism. A fourth great prophetic religion, which should be mentioned for its historic importance, is Zoroastrianism, once the national faith of the Persian Empire. Zoroaster (Zarathushtra), a prophetic reformer of *c.* 7th century BC, apparently professed a monotheistic faith and a stern devotion to truth and righteousness. At the age of 30 he experienced a revelation from Ahura Mazdâ (The Wise) and chose to follow him in the battle against the forces of evil. This revelation enabled Zoroaster and his followers to comprehend the difference between good (Truth) and evil (The Lie) and to know the one true God. Later forms of Zoroastrianism apparently had an impact on Judaism, from the time of the Babylonian Exile, and, through Judaism, on Christianity.

THEMES AND FUNCTIONS

Recurrent questions concerning revelation include the relationship between general and special revelation; the relationship between word and deed as media of special revelation; the authority of the sacred books; the revelatory value of tradition; the nonverbal component in revelation; the interpersonal dimension of revelation; and the relationship between faith and reason.

General revelation: the role of nature. The Eastern religions, on the whole, differ from Western religions in that they place less emphasis on a special or exclusive revelation received by a "chosen people" and rather speak of the manifestation of the Absolute through the general order of nature. There is, however, no irreconcilable opposition between general and special revelation. Vedānta Hinduism and Buddhism, even if they do not speak of special revelation, believe that their religious books and traditions have unique value for imparting a saving knowledge of the truth. The Bible and the Qur'ān, conversely, proclaim that although God has specially manifested himself to the biblical peoples, he also makes himself known through the order of nature. The failure of some nations to acknowledge the one true God is attributed not to God's failure to disclose himself but rather to the debilitating effects of sin on the perceptive powers of man.

Special revelation: the role of history. The Western religions differ somewhat among themselves in the ways in which they understand how special revelation occurs. Some focus simply on the direct inspiration of the divinely chosen prophets. The Judeo-Christian tradition, however, characteristically looks upon the prophets as witnesses and interpreters of what God is doing in history. Revelation through deeds is conceived to be more fundamental than revelation through words, though the words of the prophets are regarded as necessary to clarify the meaning of the events. Since the Old Testament term for "word" (*davar*) signifies also "deed" or "thing," there is no clear line of demarcation between word-revelation and deed-revelation in the Bible. The biblical authors look upon the national fortunes of Israel as revelations of God's merciful love, his fidelity to his promises, his unfailing power, his exacting justice, and his readiness to forgive the penitent sinner. The full disclosure of the meaning of history, for many of the biblical writers, will occur only at the end of time, when revelation will be given to all peoples in full clarity. The Judeo-Christian notion of history as progressive revelation has given rise to a variety of theological interpretations of world history, from St. Augustine (AD 354–430) to G.W.F. Hegel (1770–1831) and other modern thinkers.

Revelation and sacred scriptures. In those religions that look for guidance to the ancient past, great importance is attached to sacred books. Theravāda Buddhism, while it professes no doctrine of inspiration, has drawn up a strict canon (standard or authoritative scriptures)—the "Pāli canon"—in order to keep alive what is believed to be the most original and reliable traditions concerning the Buddha. Mahāyāna Buddhism, while it has no such strict canon, considers that all its adherents must accept the authority of the *sūtras* (basic teachings written in aphorisms). Zen Buddhism, in many ways the broadest development of Mahāyāna thought, sometimes goes to the point of rejecting any such written authority. Many religions view their holy books as inspired and inerrant. According to a very ancient Hindu tradition, the sages of old composed the Vedas by means of an impersonal type of inspiration through cosmic vibrations. Judaism, on the other hand, looks upon the Bible as divinely inspired. The idea of verbal dictation from God, which occurs here and there in the Bible, was applied by some rabbis to the Pentateuch, which was believed to have been written by Moses under verbal inspiration, and even to the whole Bible. Christianity, which generally accepts both the Old and New Testaments as in some sense inspired, has at times countenanced theories of verbal dictation. According to the Mormons, the Book of Mormon was composed in heaven and delivered on tablets of gold to Joseph Smith. Islām holds that the Qur'ān, an eternal heavenly book, was dictated verbatim to Muḥammad. The Prophet's companions testified that he would often turn red or livid, sweat profusely, and fall into trances while receiving revelations.

The importance of canons of scriptures

Revelation and tradition. The great religions frequently make a distinction between those scriptures that contain the initial revelation and others, at the outer fringe of the canon, that contain authoritative commentaries. In Hinduism, the four Vedas and three other ancient collections—the *Brāhmaṇas, Āraṇyakas,* and *Upaniṣads*—are *Śruti* ("that which has been heard"; *i.e.,* constitutive revelation); the other sacred writings (the *sūtras,* the law-books, *Purāṇas,* and the *Bhagavadgītā* and the *Rāmāyaṇa,* the two great epics) are *Smṛti* ("that which has been remembered"; *i.e.,* tradition). Later Judaism, while recognizing the unique place of the Bible as the written source of revelation, accords equal authority to the Talmud as traditional commentary. Among Christians, Roman Catholics and the Eastern Orthodox believe that revelation is to be found not only in the Bible but also, by equal right, in the apostolic tradition. Protestants emphasize the objective sufficiency of Scripture as a source of revelation, but many Protestants today are careful to add that Scripture must always be read in the light of church tradition in order that its true message be rightly understood. Islām holds that the Qur'ān alone contains revelation in the strict sense (*waḥy*), but it accepts tradition (*Ḥadīth*) as a supplementary source of Islāmic law. Special significance is attached to the practice (*sunnah*) of the Prophet himself and to the traditions handed down by his immediate companions.

Apostolic tradition

Revelation and experience. In most religions nonverbal communication plays an important part in the transmission of revelation. This can occur in art (notably in icons, statues, and idols), in sacred music, in the liturgy, and in popular dramas, such as the mystery plays common in medieval Europe or those still performed in Indian villages. For a deeper initiation into the revelation, it is believed

necessary to live under the tutelage of a guru (teacher), monk, or holy man. To the extent that revelation is identified with a profound and transforming personal experience, the spiritual preparation of the subject by prayer and asceticism is stressed. Among the great living religions of the world, there is wide agreement that revelation cannot be fully communicated by books and sermons but only by an ineffable, suprarational experience. In Hinduism the *Upaniṣad*s emphasize the hiddenness of God. Leaving behind all created analogies, the adept is led to the point where he comes to praise God in an adoring silence more exalted than speech. Buddhism of the Mahāyāna, especially its Zen varieties, likewise advocates ecstatic contemplation.

The Eastern mystics are here in close agreement with the Jewish Ḥasidim (mystical pietists), with the Islāmic Ṣūfīs, and with the great Christian mystics, such as Pseudo-Dionysius, the Areopagite, Meister Eckehart, and St. John of the Cross. Many theologians within Judaism (*e.g.*, Maimonides) and Eastern Christianity (*e.g.*, St. John Chrysostom, St. John of Damascus) have contended that God is best known through a negative, or "apophatic," theology that makes no positive statements about God. This idea, never absent from the medieval scholastic (intellectualist) tradition, was newly emphasized by Martin Luther, who insisted that the revealed God (*Deus revelatus*) remains the hidden God (*Deus absconditus*), before whom man must stand in reverent awe. Contemporary Roman Catholic theologians, such as Karl Rahner, maintain that even in heaven God will not cease to be, for man's finite mind, an unfathomable mystery. Revelation makes man constantly more aware of the depths of the divine incomprehensibility.

The revealed God as the hidden God

Revelatory relationships. In certain forms of mysticism, particularly prevalent in the Eastern religions, the envisioned goal is an absorption into the divine, involving the loss of individual consciousness. In the Western religions and in Bhakti Hinduism the abiding distinctness of the individual personality is affirmed. Islāmic orthodoxy, looking upon revelation as a declaration of the divine will, stresses not so much the communion of man with God as rather man's obedient submission to the creator. Islāmic Ṣūfism, however, resembles Ḥasidic Judaism and Christianity in its aspiration for personal union with God. For many contemporary religious thinkers, such as the Jewish philosopher Martin Buber and the Roman Catholic philosopher Gabriel Marcel, revelation involves a mutual self-giving of the revealer and the believer in personal intercommunion. According to Karl Rahner, revelation consists primarily and essentially in God's gracious communication of his own divine life to man as a personal spirit. In his view, the articulation of revelation in the scriptures and creeds is a secondary stage, presupposing an experiential encounter with the divine. This secondary phase, however, is viewed as necessary in order that man may realize himself in his humanity as a believer and achieve solidarity with his fellow believers. In general, the Western religions tend to attach more importance to the idea of a community of faith than do the Eastern religions. Revelation in the biblical and Islāmic view is addressed not to individuals as such but to a whole people, which achieves its identity, in part, by articulating its faith in writings that are approved as authentic expressions of what God has revealed.

Revelation and reason. The problem of the relationship between revelation and reason arises, on the one hand, because revelation transcends the categories of ordinary rational thought and, on the other hand, because revelation is commonly transmitted by means of authoritative records, the contents of which cannot be verified by the believer. Buddhism, since it does not attribute inspiration or inerrancy to its canonical sources, allows some scope for individual reason to criticize the authoritative writings, but, like other religions, it has to face the charge that the illumination to which it aspires may be illusory. Orthodox Hindus, giving full authority to the Veda, hold that human reason errs whenever, on the grounds of perceptual experience, it takes issue with the sacred writings. Hinduism, however, allows for great freedom in the exegesis (interpretation) of its sacred books, some of which are more poetic than doctrinal.

The tension between faith and reason has been particularly acute in the Western religions, which find revelation not simply in holy books but in prophetic words that call for definite assent and frequently command a precise course of action. The ambiguities of scripture in these religions are frequently cleared up by creeds and dogmas of the community, calling for the assent of true believers. Judaism, Christianity, and Islām, moreover, came into close contact with Hellenistic culture, which held up the ideal of rationally certified knowledge as the basis for the good life. They, therefore, had to face the problem: could assent to an authoritative revelation be justified before the bar of reason? Some theologians took a "fideist" (faith-based) position, maintaining that reason must in all things submit to the demands of revelation. Others, such as the Arabic philosopher Averroës and his followers (both Muslim and Christian), accepted the primacy of reason. They reinterpreted the content of revelation so as to bring it into line with science and philosophy. A third school, in which the medieval Jewish philosopher Maimonides and the medieval Christian scholastic theologian Thomas Aquinas may be included, sought to maintain the primacy of faith without sacrificing the dignity of reason. According to the Thomist theory, human reason can discern the credibility of revelation because of the external signs by which God has authenticated it (especially prophecies and miracles). Reason, moreover, makes it possible for the believer to understand, in some measure, the revealed mysteries. This intellectualist position continues to appeal to many Christians; but some maintain that it overlooks the qualitative differences between faith—as a transrational assent to mystery—and scientific knowledge, which operates within the categories of objectivizing reason.

Tension between faith and reason

CONCLUSION

In some theological circles the concept of revelation is rejected on the ground that it is bound up with mythological and anthropomorphic conceptions and introduces an unassimilable element into the history of religions. It would seem, however, that the concept can be purified of these mythical elements and still be usefully employed. In the sphere of religion, wisdom is often best sought through privileged moments of ecstatic experience and through the testimony of those who have perceived the sacred or holy with unusual purity and power. The self-disclosure of the divine through extraordinary experiences and symbols is fittingly called revelation. Because of the pervasiveness of the idea of revelation in the world's religions and because the various religions have had to cope with similar theological problems concerning revealed knowledge, revelation has become a primary theme for dialogue among the great religions of mankind.

(A.Du.)

Perception of the sacred or holy

Covenant

The concept of covenant—that is, a binding promise—is of far-reaching importance in the relations between individuals, groups, and nations. It has social, legal, religious, and other aspects. This section is concerned primarily with the term in its special religious sense and especially with its role in Judaism and Christianity.

NATURE AND SIGNIFICANCE

Covenants in the ancient world were solemn agreements by which societies attempted to regularize the behaviour of both individuals and social organizations, particularly in those contexts in which social control was either inadequate or nonexistent. Though ancient pre-Greek civilizations apparently never developed a descriptive theory of covenants, analysis of covenant forms and the ancient use of language yields a definition that essentially is the same as that found in modern law. It is a promise or agreement under consideration, usually under seal or guarantee between two parties, and the seal or symbol of guarantee is that which distinguishes covenant from modern contract.

Definition of covenant

The concept of covenant has been of enormous importance in the biblical tradition; from it there is derived the

long traditional division of the Bible into the Old and New Testaments (Covenants). In postbiblical Judaism and sporadically in Christianity, the concept of covenant has been a major source and foundation of religious thought and especially of the concept of the religious community, but the nature and content of covenant ideas have undergone an extremely complex history of change, adaptation, and elaboration.

Though both covenant and law in the ancient world were means by which obligation was both established and sanctioned, and are often virtually identified with each other in modern scholarly literature, there are, nevertheless, very important contrasts between the two that should not be obscured. A covenant is a promise that is sanctioned by an oath. This promise in turn was accompanied by an appeal to a deity or deities to "see" or "watch over" the behaviour of the one who has sworn, and to punish any violation of the covenant by bringing into action the curses stipulated or implied in the swearing of the oath. Legal procedure, on the other hand, may be entirely secular, for law characteristically does not require that each member of the legal community voluntarily swear an oath to obey the law. Further, in ordinary legal procedure the sanctions of the law are carried out by appropriate agencies of the society itself, not by transcendent powers beyond the control of man and society.

Because a person can bind only himself by an oath, covenants in the ancient world were usually unilateral. In circumstances in which it was desirable to establish a parity (equivalence) treaty, such as in rare cases in political life, the parity was obtained by the simple device of what might be termed a double covenant, in which both parties would bind themselves to identical obligations, and neither was therefore subjected to the other.

The oath was usually accompanied by a ritual or symbolic act that might take any of an enormous range of forms. One of the most frequent of these was the ritual identification of the promisor with a sacrificial animal, so that the slaughter and perhaps dismemberment of the animal dramatized the fate of the promisor if he were to violate the covenant.

ORIGIN AND FUNCTION OF COVENANTS

Origins. That covenants most probably originated in remote prehistoric times is indicated by the fact that they were already well-developed political instruments by the 3rd millennium BC. To judge from later parallels and from the modern observations of anthropologists, covenants may very well have developed at least in part out of marriage contracts between exogamous tribes or bands; *i.e.*, those groups that stayed within the required patterns of intermarriage. Whether or not this was the case, the most important functions of covenants for 1,000 years before the 13th-century BC Sinai covenant (see below) had to do with the creation of new relationships, both familial and political. Though the old theory of "social contract"—*i.e.*, the basic agreement about the social and political order—as the basis of large social organizations has not for some time been much in favour among social scientists, very early historical evidence seems increasingly to suggest that covenants may have been much more instrumental in society than has been realized.

Early functions of covenants

Typically, so far as existing sources now reveal, a covenant between social groups regularized in advance the relationships between two societies after one had been subjugated by a superior coercive force, usually by military action or the threat thereof. In the Mari documents (18th-century-BC archives from the palace at Mari in Syria), such a covenant was called a *salimum*, a "peace," probably because the promises made by the vanquished brought to an end the necessity of military operations against the vassal ruler or state. As is the case throughout so much of human history, ancient states characteristically seem to have regarded their neighbours as either enemies or vassals. Thus it is not surprising that covenants made under duress had little vitality, particularly when the terms of the covenant called for a considerable annual tribute to the overlord state.

Late Bronze Age developments. About the beginning of the late Bronze Age (*c.* 1500 BC), there occurred a major step forward in both the form and the concept of political covenants as is attested by treaties of the Hittite Empire of Asia Minor. Though the realities of political life were probably little changed, since the foreign policy of the Hittite Empire was primarily military, the structure of suzerainty treaties from this time on included rather strenuous efforts to demonstrate that the vassal's obligations to the Hittite overlord were really founded upon the former's self-interest, not merely upon the brute military force of the latter.

By far the most evidence for international treaties in the ancient world comes from Hittite sources, which were contemporary with the events that preceded and led up to the formation of the ancient Israelite federation of tribes in Palestine. The treaty form in written texts was highly developed and flexible but usually exhibited the following structure: preamble, historical prologue, stipulations, provisions for deposit and public reading, witnesses, and curses and blessings formulas. (1) The preamble names the overlord who grants the treaty-covenant to the vassal. The titles and laudatory epithets of the Great King are also given. (2) The historical prologue describes the previous relationships between the two parties in some detail, usually emphasizing the benevolent acts of the Great King toward the vassal. Thus the covenant is based upon the demonstrated benefits that have already been received and therefore holds out the expectation of continuing advantage for faithful obedience to the covenant. There is an implication that obedience to obligations is based upon gratitude. (3) The stipulations, which in form are much like those of the ancient Mesopotamian law codes (case law), define in advance the obligation of the vassal in certain circumstances. In addition, there are also generalized statements of obligation of a type that has been called "apodictic law" (regulations in the form of a command). The obligations deal particularly with military assistance, the treatment of fugitives, and foreign policy. Treaty relationships with other independent states are a violation of covenant. (4) Provision is made for deposit of the treaty in the temple and for periodic public reading. Because the temple is the "house of the god," the written document was placed there for the watchful attention of the deity. The treaty obligations, however, were also binding upon the vassal's citizenry, and so at stipulated intervals the text was read to the assembly, both as a reminder and a warning. (5) The list of witnesses included, in addition to the major deities of both states, deified elements of the natural world, such as mountains, rivers, heaven and earth, winds, and clouds. The witnesses were those powers that were believed to be beyond human control and upon which man and society were regarded as completely dependent. They were invoked to apply the appropriate sanctions of the covenant. (6) The curses and blessings formulas are the sanctions that furnish not only negative but also positive motivations for obedience. They include the natural and historical calamities beyond human control, such as disease, famine, death without posterity, and destruction of the society itself. The blessings are of course the opposite: prosperity, peace, long life, continuity of kingship and society.

Hittite treaties

In view of the obsession with rituals that characterized Hittite culture, some elaborate ceremony probably accompanied the ratification of covenant, such as the account of one preserved in the document known as "The Soldiers' Oath," but it is not described in existing covenant texts.

Scholars in Europe and America in the 20th century have seen an astounding similarity between this treaty structure and the biblical traditions of the Sinai covenant. Publication of texts in the mid-1950s was followed by an enormous amount of scholarly discussion, but as yet no conclusions can be said to represent a scholarly consensus. The formal similarity to biblical traditions cannot be denied, but the problem of what historical conclusions can be drawn from the formal similarities is highly sensitive and controversial. While the following synthesis is a probable, and historically plausible, interpretation, it must be admitted that other possibilities can by no means be excluded.

THE ORIGIN AND DEVELOPMENT OF BIBLICAL COVENANTS: JUDAISM

The Sinai covenant. *Historical background.* The 100 years between 1250 and 1150 BC saw the complete destruction, or reduction to virtual impotence, of every major political state in the eastern Mediterranean region and the beginning of a "dark age" that has yielded very few written materials from which historical conclusions can be drawn. The reasons for the universal catastrophe are far from clear, but the reversion of society to communities of peasants and shepherds with a subsistence level economy can be well illustrated archaeologically. The earliest biblical traditions illustrate the conditions in Palestine at this time, though it is a difficult task to distinguish genuine ancient traditions from the use of the past by biblical writers to give religious validity to social realities or institutions of much later date.

In view of the highly elaborate social structure of the old Bronze Age states—with its apex in the military aristocracy, a highly complex priesthood, and ritual—and the equally complex social structure of the many local enclaves and tribes—each with its particular god—the monotheistic and ethically centred religious ideology of early Israel has been regarded for millennia as a miracle of "revelation," which cannot be explained on the basis of usual historical principles and concepts. Yet, ancient Israel was an historically existent community created, and precariously maintained, by a unity of which the religious ideology was the foundation for two centuries, until military considerations resulted in the formation of a political centralization of power about 1000 BC. The covenant tradition is the only instrument by which the effective functioning of that unity can be understood, and its importance is underlined by the biblical traditions themselves. The structure of the Hittite treaties now makes available an historical precedent that enables scholars to understand the structure of early Israelite thought and consequently its functional operation in history.

The covenant tradition as an instrument of unity

The covenant at Sinai. The Decalogue (The Ten Commandments) given by Yahweh, the God of the Israelites, at Sinai, plus the various traditions associated with earliest Israel yield all of the important elements of the Hittite treaty form but in an extremely succinct and simple form. Yahweh is identified as the covenant giver, and the historical prologue is the only possible one according to the ancient traditions: the announcement that it is this God who delivered the assembled group from bondage in Egypt (in the 13th century BC). This delivery is a free, voluntary act of the deity that forms the basis of the obligations that the community can either accept together with a lasting relationship to that God or reject, thus entailing a permanent hostility (hatred) between the God and human beings. It is the common relationship to a single sovereign God that furnished the basis for a radically new kind of community, which grew with rapidity first in Transjordan, then in Palestine proper, until it included virtually all the nonurban population of the region.

The new community was the answer, temporarily at least, to the old dilemma of civilization: how to maintain peace among a large and diverse population, perform the necessary social functions of cooperation and protection, and control individual attacks upon the security and property of others without the enormous and expensive paraphernalia of political bureaucracy, military machine, and the ruinous tax collector. It was, for all functional purposes, the Kingdom (or Rule) of Yahweh, which excluded the deification of any other factor in human history or nature that was of importance to human life and well-being. The Sinai covenant marked the beginnings of nearly all the various theological themes that were to be so greatly elaborated upon in the following millennia: the Providence, or Grace, of God; the Kingdom of God; the sin of man and the wrath of God; the Holy People as the community of God; the rewards and punishments of the obedient and the disobedient respectively; and above all, the ethical norms as the essence of divine command over against the universal pagan obsession with proper ritual as the normative expression of man's subjection to the divine will.

The Sinaitic covenant stipulations may be expressed in modern functional terms in the following manner: (1) The commandment to have no other gods involves the obligation to refuse subjection to all other social and human concerns and their symbolization in art forms so as to give them a position of parity or superiority to Yahweh and his commands. (2) The commandment not to take the name of God in vain emphasizes the unconditional sanctity of oaths that Yahweh was called upon to guarantee and enforce. (3) The commandment to observe the Sabbath, the seventh day, the original social function of which is still unknown, could very well have grown out of a common village custom, for even in Rome in the 1st century BC, good farming practice permitted work animals and slaves to rest every eighth day—and this is precisely the interpretation given in Deut. 5:14. (4) The commandment to honour father and mother emphasizes the treatment of parents with respect and deference, which must have been of particular importance in a time of social upheaval and polarization. (5) The commandment not to kill meant that killing of persons by persons, even by accident if it involved negligence, was a usurpation of the divine sovereignty over persons. Contrary to modern reinterpretations, among opponents of capital punishment and pacifists, this could not include execution of persons for crime or killing of the enemy in warfare, for in both cases human beings were acting as the agents of Yahweh under divine command, just as the various officials of states have long carried out similar functions without incurring personal guilt for their acts. (6) Other commandments against theft, adultery, and false witness categorically prohibit acts that call into question the security of property, of family relationships and true lineal succession, and the integrity and therefore the justice of juridical procedures in society. (7) Finally, the prohibition of coveting what one's neighbour has excludes an enormous range of social attitudes and motivations that modern man now takes for granted as normal, if not essential.

Modern functional equivalents of the Ten Commandments

Most, if not all, of the Ten Commandments are ethical obligations of which violations are very difficult if not impossible for society to detect, much less to enforce or punish. The Sinai covenant, therefore, marked the beginnings of a systematic recognition that the well-being of a community cannot be based merely upon socially organized force, nor can the political power structure be regarded, as in ancient pagan states, as the manifestation of the divine, transcendent order of the universe.

The Ten Commandments as ethical obligations

Post-Sinai covenants. Traces remain in the biblical traditions to indicate that the new community formed from a "rabble" at Sinai was in very short time joined by a considerable part of the population of Transjordan and Palestine proper. After the destruction (in the late 13th century) of the military chiefdoms ruled by Sihon and Og in the area east of the Jordan River, the Hebrews held a covenant ceremony at Shittim (northeast of the Dead Sea), which has been greatly elaborated upon in tradition as the "second giving of the Law," Deuteronomy. Though it is true that the Book of Deuteronomy from the 7th century BC exhibits the same basic structure as that of the old covenant form, it is at present impossible to reconstruct the original form or content of the Shittim covenant. It may be presumed that entry into the community by covenant was followed by the allotment of land as tenured fiefs from Yahweh and the organization of the population into "tribes." This organization probably was the last event of the Hebrew leader Moses' life, and the sequel in the more important covenant at Shechem (northwest of the Dead Sea) took place under the leadership of Joshua, the successor of Moses.

Shechem evidently had had an important covenant tradition long before Israel existed. The name of its god, Baal Berit ("Lord of the Covenant"), presupposes some kind of covenant basis for the local social structure, just as a considerable segment of the population can be shown to have originated from Anatolia.

The Shechem covenant narrative has been preserved at least in part in Joshua, in which Joshua appeals to the family and clan heads to choose between the new dominion of Yahweh and the continuation of the old ancestral cults of the Amorite tradition "beyond the River." As

The Shechem covenant

in the case of the Transjordan covenant at Shittim, this covenant followed the defeat of a coalition of petty kings and evidently the removal of many others according to the list of Joshua. Again, there ensued an allotment of fields and an organization of the population into administrative units called "tribes," each under a *nasi* (literally, "one lifted up").

The entire process from the covenant at Sinai to the unification of perhaps a quarter of a million people by a covenant involving a religious loyalty to a single deity took only a little over one generation. It began with a group of probably considerably less than 1,000 people who left Egypt with Moses.

The subsequent history of the Sinai covenant tradition is very complex. The Book of Deuteronomy preserves slight traces of a covenant-renewal ceremony held every seven years, which is inherently plausible and which would function as a means for obtaining the oath-bound loyalty to Yahweh and his dominion of those who had come into the community from the outside or who had come of age in the intervening period.

The covenants of the Israelite monarchy (1020–587/586 BC). Since early Israel was a religious confederacy of tribes that bitterly rejected the old military chiefdoms and their religious ideology, which elevated a Baal, or local agricultural deity—the god with the club as a symbol of the supernatural power undergirding the king—to a position of preeminence in the pantheon, it follows that the authentic Yahwist traditions stemming from Moses could not furnish a religious ideology to legitimize the monarchy when it was finally established first under Saul (reigned *c.* 1020–1000 BC) and then successfully under David (reigned *c.* 1000–962). Furthermore, early in David's reign, he had incorporated by military force most of the existing city-states of Palestine and Transjordan into his empire, and that population had never given up the old Bronze Age cults.

It is not surprising, therefore, that this double dilemma of the new political structure should have driven the royal bureaucracy to pre-Mosaic sources as a solution to the problem. One result was the reintroduction of the age-old pagan concept of the king as the "chosen" one of the gods and a radically different—and opposite—concept of covenant, in which it was now Yahweh, not the king or the people, who bound himself by oath. Possibly modelled after old royal covenants by which ancient pagan kings made a grant to their faithful retainers, the Davidic covenant introduced a radically different (and thoroughly pagan) element into the Mosaic tradition, and the two traditions contended with one another for the next 1,000 years.

Since the old Israel-Jacob (pre-Mosaic) traditions also could not furnish an ideological base for unifying the old Israelite and non-Israelite populations under the monarchy, pre-Mosaic epic traditions of Abraham (perhaps 19th–18th centuries BC) were appealed to to furnish the "common ancestor" symbol of unity, and the covenant tradition—no doubt, already a part of that epic—was readapted to bring it up to date. The deity (now identified with Yahweh) bound himself by oath to fulfill certain promises to Abraham, though the content of the promise, in the form now received, was by and large a description of the historical situation of the Davidic empire. Though it is difficult to see what the social or ideological function may have been, the covenant with Noah (the hero of the Flood) in Genesis exhibits the same structure. The result of all these radical changes in a very short time was a complete confusing of the religious tradition and structure and a permanent deposit of the pre-Mosaic pagan religious ideology into the biblical tradition. It seems virtually certain that the Sinai tradition was itself systematically reinterpreted in the so-called ritual decalogue of Exodus in which it is dogmatically stated that the Sinai obligations were entirely ritual in nature, rather than ethical-functional. The first tables of stone of the Ten Commandments, after all, had been "broken," which in the ancient world was a customary phrase used to indicate the invalidation of binding legal documents.

Reinterpretation of the Decalogue

The next several centuries illustrate the constant battle between the Mosaic and the reintroduced pagan elements. The prophets proclaimed and supported the disintegration (*c.* 922 BC) of the Solomonic empire into a northern (Israel) and a southern (Judah) kingdom as the divine chastisement of Yahweh for gross disobedience. Particularly in the north, which did not retain the Davidic dynasty, the prophets periodically proclaimed the necessity and inevitability of wiping out one royal dynasty after another. Elijah, a 9th-century BC rustic prophet, ridiculed the idea that the Israelites could limp along on both legs—*i.e.,* observe loyalty to both the Yahwistic and the Baal cults. Reforms were carried out occasionally, but not until the time of Josiah, the young king of Judah (late 7th century BC), and the discovery of an old copy of the Mosaic legal-ethical tradition (the Deuteronomic code) in *c.* 621 BC was serious reform undertaken—and there with little permanent success. The preservation of the Mosaic tradition was a function of the destruction of the monarchical state and its religious symbol, the temple, which nearly all the pre-exilic (before 587/586 BC) prophets had predicted.

The post-Exilic covenant tradition. Though the prophet Jeremiah (late 7th century BC) had predicted a "new covenant" written upon the heart (Jeremiah), not until the time of the prophets Ezra and Nehemiah in the 5th century is there another biblical narrative of covenant making, this time one of incalculable importance for the future of both postbiblical Judaism and Christianity and perhaps even for certain aspects of political theory or practice in the West (*e.g.,* "Covenant" of the United Nations, Mayflower Compact, and constitutions).

The account in Nehemiah is not so much that of a covenant as it is of a constitutional convention, the purpose of which was to establish as binding law the complex of traditions that had been preserved and recorded as the "law of God which was given by Moses, the servant of God" (Nehemiah). It is a one-party enactment by the authorities and representatives of the community, in which Yahweh appears only as the deity addressed in the long historical prologue in the form of a prayer. The content is a recapitulation of the Deuteronomic history (interpretations of the 7th-century BC document), narrating the benevolent acts of Yahweh and the sin and punishment of the people. In order to avoid the curses, and obtain the blessings, the community resolved henceforth to observe the "law of God." From this time on, the dominant concept of covenant in Judaism identifies it with circumcision, the ritual by which on the eighth day of his life, the male Jew becomes obligated to obey the law of Moses, the *berit* (covenant). The Sinai covenant had become permanently identified with the accumulation of legal-ritual tradition, and the community was identified not as the complex variety of all those who wished and accepted the rule of God but as the ethnic group of those who were heirs of the promise to Abraham in direct lineal (and fictitious) descent.

THE ORIGIN AND DEVELOPMENT OF THE COVENANT IN CHRISTIANITY

The New Testament tradition of the covenant. The cup of wine at the Last Supper of Jesus and his disciples before Jesus' crucifixion is identified in all New Testament sources as the (new) covenant by Jesus himself, but in spite of millennia-long controversy, theological elaboration, and discussion, the nature and meaning of the covenant has never been adequately understood historically, and the variety of interpretations regarding covenant in the New Testament itself indicates that very early in the tradition it had become a problem. Here it is possible only to indicate some significant associations that might explain why it was called a "covenant" and how the ancient Sinaitic tradition was radically renewed but the basic structure retained.

First, it has been noted that a most important aspect of covenant traditions common to most ancient cultures was the ritual identification of the oath taker with the sacrificial victim. The identification of the bread and the wine with the body and blood of Christ at the Last Supper apparently was interpreted in this sense, so that the subsequent death of the victim entails the symbolic death—the ultimate curse for breach of covenant—of all those

The death of Jesus in relation to the covenant

who were thus identified with the victim. Consequently, the curses of the law were nullified. The death of Jesus thus becomes in the Christian proclamation the centre of the historical narrative—the historical prologue of the covenant—leading up to the covenant enactment, or the *sacramentum,* to use the Latin term of the early church, which in secular use at that time meant primarily the soldier's oath of loyalty to the emperor (see above *Late Bronze Age developments*). The Christian covenant was thus a highly complex historical act that brought about a relationship of the believer to Christ whose (normally) unseen Glory was identified with that of God himself, whose Lordship was viewed as operational in history, and whose community (of believers) was identified with the Kingdom (Dominion or Rule) of God. If God in the Old Testament could rule without kings, God could, for the New Testament writers, rule without the elaborate structure of the accumulated legal traditions. They were regarded as valuable for edification and for warning but no longer as having binding validity. The anathema, or curse, was no longer tied to the definitions of legal violation but rather to rejection of God's rule in Christ. The community in turn was no longer the lineal descent group with a parochial ritual tradition but the assembly (*ekklēsia*) of those who had through the covenant accepted a relationship to the dominion of Christ.

The obligations could not, in the New Testament viewpoint, be again defined in legal terms, nor could they be enforced by social power structures, which could deal only with external formal acts, not with the basic springs of behaviour, such as love or hate. The content of obligation was thus not defined; instead, in the Sermon on the Mount (Matthew) and other New Testament literature, it is the criteria (motivations, ethical norms, personality traits) by which the rule of God is recognized upon which the emphasis falls. The presumption is that anyone who is capable of recognizing the rule of God in his experience in society will also be capable of understanding what the nature of his obligation will be in specific circumstances. The curses and blessings alike are then postponed until the final judgment. The motivations of fear of punishment and hope of reward are irrelevant to the daily routine of ethical choice, which is thus not only possible (*i.e.,* not prescribed in advance by legal definition) but unavoidable and also necessary to make responsible ethical decisions in a world that is characterized by cultural diversity and change.

The post-apostolic church. Covenant concepts in early Christian theology apparently centred on the transferrence of the Davidic covenant to the Messianic figure—*i.e.,* Christ. The fundamental theological problem of the early church was to validate the authority of Christ against both paganism and Judaism and to maintain the authority of the new religious community. After the great theologian Augustine (354–430), little attention was given to covenants until the Reformation in the 16th century. Though Luther (1483–1546) referred to and discussed the biblical covenants, it was never of particular importance to his theology. It is rather in Reformed theology, particularly that of John Calvin (1509–64) and the later Puritans of the 17th century, that its further elaboration took place. One aspect of the use of covenant may be cited in the famed Mayflower Compact of November 11, 1620 (drawn up by the Pilgrims, Separatists from the Church of England) by which a "civil body politic" was formed that would in turn enact laws and offices for the general good.

Significance of covenants in Reformed Protestant theology

The theological elaboration of covenant in Puritan and Separatist theology centred on the themes of election, grace, and Baptism. It is curiously ironic that covenant enactment, such as the Mayflower Compact, became historically operative but remained essentially secular, while the religious covenant became predominantly a theological concept associated particularly with Baptism—the ritual means by which a person became a participant in the covenant of grace. The essential elements in the biblical covenant—*i.e.,* that of free, voluntary acceptance of ethical obligation on the basis of and as response to past experience—has virtually always given way to covenant as fixed religious dogma that legitimizes the social structure. Covenant historically has been a means by which new communities are formed, particularly in times of rapid change, social dislocation, or political breakdown. Covenants have rarely been the actual instruments by which societies actually functioned for long, but they are extremely frequent as ideological foundations for sociopolitical legitimacy.

COVENANT IN OTHER RELIGIONS

Islām. Covenants (*mithāq, 'ahd*) were of great importance in the formative period of Islām (7th century AD, or 1st century AH—after the Hegira, Muḥammad's flight from Mecca to Medina). More than 700 verses of the Qur'ān, the Muslim sacred scripture, have to do with various aspects of covenant relationships. As one recent Muslim writer, Sayyid Qutb, states, Islām combines both the Old and the New Testaments (covenants) and the Last Covenant, of Islām, as well. All revelation from Adam to Muḥammad is regarded by Muslims as a unit, mediated through a series of prophets, or messengers, with whom God made a covenant: Noah, Abraham, Moses, and Jesus. Though the concept is difficult, it seems that the prophet in each case was given a revelation and a religion to which he covenanted with God to witness faithfully. This concept of a covenant of the prophets conveys the conviction of the unity of revelation as well as the unity of God in past history.

On the second level, the Muslim community itself is often regarded as being composed of those who have accepted the covenant with God. In this connection, the grace, or providence, of God in nature or creation is of great importance. In addition to this view is the repeated emphasis upon the doctrine that God alone is man's sole benefactor, and for these reasons the response of gratitude is an important element in the structure of the covenant. It is also necessary that rewards and punishments are included. These are predominantly, as in the Christian concepts, focussed upon the hereafter, paradise, and hell, though not exclusively so. The recipients of the rewards and punishments are described as those who obey or disobey Allāh's (God's) commands, which include prayer, paying the *zakāt* (head tax: an obligatory charity), belief in the messengers of Allāh, fearing God alone, and refraining from theft, adultery, murder, and false witness. They are further obligated to show kindness to parents and to strive in the cause of God with their persons and property.

Importance of the covenant in the formative period of Islām

On the historical and social level, it seems quite certain that the community of the formative period in Islām was based on covenant acts, in which persons or groups formally proclaimed their acceptance of Muḥammad's message and swore an oath of loyalty, accepting the obligations outlined above. References to the clasp of hands indicate that this was probably regarded as the formal act of commitment and acceptance by the community. In later Muslim theology, as in Christianity, the covenant idea seems to have been of comparatively little importance.

Other religions. It seems that only in the religions stemming from the biblical tradition is covenant of central importance. Though gods are often invoked as guarantors of promises sworn to in Iranian and Indic (Hindu) religious traditions, the covenant with a deity or the community as a covenant-bound one apparently was of relatively little importance, or possibly the concept has not been recovered by modern scholarship. The great importance of Mithra in early Iranian religion as god of the covenant and Mitrā-Varuṇa in Indic (Hindu) religion suggests that such concepts may have been more important than is now realized. Thus, modern scholarship has yet to indicate the importance of the covenant concept in Indo-Iranian and other religions. (G.E.Me.)

Prophecy

Prophecy, a religious phenomenon generally associated with Judaism and Christianity, is found throughout the religions of the world, both ancient and modern. In its narrower sense, the term prophet (Greek *prophētēs,* "forthteller") refers to an inspired person who believes that he has been sent by his god with a message to tell. He

is, in this sense, the mouthpiece of his god. In a broader sense, the word can refer to anybody who utters the will of a deity, often ascertained through visions, dreams, or the casting of lots; the will of the deity also might be spoken in a liturgical setting. The prophet, thus, is often associated with the priest, the shaman (a religious figure in primitive societies who functions as a healer, diviner, and possessor of psychic powers), the diviner (foreteller), and the mystic.

In a much broader sense, the term prophet has been used in connection with social and religio-political reformers and leaders.

NATURE AND SIGNIFICANCE

A primary characteristic of prophetic self-consciousness is an awareness of a call, which is regarded as the prophet's legitimization. This call is viewed as ultimately coming from a deity and by means of a dream, a vision, an audition, or through the mediation of another prophet. The Old Testament prophet Jeremiah's call was in the form of a vision, in which Yahweh (the God of Israel) told him that he had already been chosen to be a prophet before he was born (Jer. 1:5). When the call of the deity is mediated through a prophet who is the master of a prophetic group or an individual follower, such a call can be seen as a mandate. Furthermore, such mediation means that the spirit of the prophet master has been transferred simultaneously to the disciple. In the case of cult prophets, such as the prophets of the gods Baal and Yahweh in ancient Canaan, the call may be regarded as a mandate of the cult.

Prophets were often organized into guilds in which they received their training. The guilds were led by a prophet master, and their members could be distinguished from other members of their society by their garb (such as a special mantle) or by physical marks or grooming (such as baldness, a mark on the forehead, or scars of self-laceration).

The nature of prophecy is twofold: either inspired (by visions or revelatory auditions), or acquired (by learning certain techniques). In many cases both aspects are present. The goal of learning certain prophetic techniques is to reach an ecstatic state in which revelations can be received. That state might be reached through the use of music, dancing, drums, violent bodily movement, and self-laceration. The ecstatic prophet is regarded as being filled with the divine spirit, and in this state the deity speaks through him. Ecstatic oracles, therefore, are generally delivered by the prophet in the first-person singular pronoun and are spoken in a short, rhythmic style.

Prophetic ecstasy

That prophets employing ecstatic techniques have been called madmen is accounted for by descriptions of their loss of control over themselves when they are "possessed" by the deity. Prophets in ecstatic trances often have experienced sensations of corporeal transmigration (such as the 6th-century-BC Old Testment prophet Ezekiel and the 6th–7th-century-AD founder of Islām, Muḥammad). Such prophets are believed to have a predisposition for such unusual sensations.

The functions of the prophet and priest occasionally overlap, for priests sometimes fulfill a prophetic function by uttering an oracle of a deity. Such an oracle often serves as part of a liturgy, as when ministers or priests in modern Christian churches read scriptural texts that begin with the proclamation: "Thus says the Lord." The priest, in this instance, fulfills the prophetic function of the cult. Not only do the roles of the prophet and priest overlap but so do the roles of the prophet and shaman. A shaman seldom remembers the message he has delivered when possessed, whereas the prophet always remembers what has happened to him and what he "heard."

The diviner, sometimes compared with the prophet, performs the priestly art of foretelling. His art is to augur the future on the basis of hidden knowledge discerned almost anywhere, as in the constellations (astrology), the flight of birds (auspices), in the entrails of sacrificial animals (haruspicy), in hands (chiromancy), in casting lots (cleromancy), in the flames of burning sacrifices (pyromancy), and other such areas of special knowledge (see also OCCULTISM: *Divination*; SACRED OFFICES AND ORDERS: *Shamanism*).

Mystics and prophets are similar in nature in that they both claim a special intimacy with the deity. The mystic, however, strives for a union with the deity, who usurps control of his ego, whereas the prophet never loses control of his ego. On occasion mystics have delivered messages from the deity, thus acting in the role of a prophet, and have been known to use ecstatic trances to reach the divine or sacred world; *e.g.*, many Roman Catholic saints and Muslim Ṣūfīs (see below *Saint;* see also RELIGIOUS EXPERIENCE: *Mysticism*).

In the Western world, Israelite prophecy is regarded as unique, for not only did it oppose institutionalized religion but it is understood as having propagated an ethical religion emphasizing individual freedom, a religion not dependent on mechanical ritual and legalism.

Uniqueness of Israelite prophecy

The term prophecy also has been used in a strictly predictive sense, not necessarily dealing with religious themes. In this sense, *The Communist Manifesto,* by Karl Marx and Friedrich Engels, was viewed as a "prophecy" of things to come; a new approach that goes against the traditional in literature, art, politics, and other areas may—in this wider sense—be termed "prophetic."

TYPES OF PROPHECY

Types of prophecy can be classified on the basis of inspiration, behaviour, and office. Divinatory prophets include seers, oracle givers, soothsayers, and mantics (diviners), all of whom predict the future or tell the divine will in oracular statements by means of instruments, dreams, telepathy, clairvoyance, or visions received in the frenzied state of ecstasy. Predictions and foretellings, however, may also be the result of inspiration, or of common sense by the intelligent observation of situations and events, albeit interpreted from a religious point of view.

Of broad importance to the religious community is the cult prophet, or priest-prophet. Under the mandate of the cult, the priest-prophet (who may be an ordinary priest) is part of the priestly staff of a sanctuary, and his duty is to pronounce the divine oracular word at the appropriate point in a liturgy. As such, he is an "institutional" prophet. The difference between a cult prophet and a prophet in the classical sense is that the latter has always experienced a divine call, whereas the cult prophet, pronouncing the word of the deity under cultic mandate, repeats his messages at a special moment in the ritual. Because of the timeless character of cultic activity, however, every time he prophesies, his message is regarded as new.

Missionary (or apostolic) prophets are those who maintain that the religious truth revealed to them is unique to themselves alone. Such prophets acquire a following of disciples who accept that their teachings reveal the true religion. The result of this kind of prophetic action may lead to a new religion; *e.g.,* Zoroaster, Jesus, and Muḥammad. The founders of many modern religious sects also should be included in this category.

The role of prophecy in the founding of new religions

Another type of prophet is of the reformative or revolutionary kind (looking to the past and the future), closely related to the restorative or purificatory type (looking to the past as the ideal). The best examples are the Old Testament classical prophets; *e.g.,* Amos and Jeremiah. Many of these so-called literary prophets were working to reform the religion of Yahweh, attempting to free it from its Canaanite heritage and accretions. In the Arab world Muḥammad is included in this category. The social sympathy found among such prophets is rooted in their religious conscience. What may have been preached as religious reform, therefore, often took on the form of social reform. This kind of prophecy is also found in India and Africa, where prophets in modern times have arisen to restore or purify the old tribal religious forms, as well as the customs and laws that had their sources in the older precolonial religious life. Many of these movements became revolutionary not only by force of logic but also by force of social and political pressure (see above, *Eschatology*).

Though there may be several categories of prophecy according to scholars, no sharp line of demarcation differentiates among these different types. Any given prophet may be both predictive and missionary, ecstatic as well as reformative.

PROPHECY IN THE ANCIENT MIDDLE EAST AND ISRAEL

The ancient Middle East. In ancient Egypt, charismatic prophecy apparently was not commonplace, if it occurred at all, though institutional prophecy was of the greatest importance because life was regarded as depending upon what the gods said. Some ancient texts contain what has sometimes been regarded as prophetic utterances, but these more often are considered to be the product of wise men who were well acquainted with Egyptian traditions and history. Among Egyptian sages, historical events were thought to follow a pattern, which could be observed and the laws of which could be discerned. Thus, times of hardship were always thought to be followed by times of prosperity, and predictions were made accordingly.

Egyptian texts

In Egyptian mantic (divinatory) texts there are prophetic sayings, but the particular concerns of these texts are more political than religious. Some are fictitious, and many are considered to have been prophesied after the event has already taken place. The papyrus text "The Protests of the Eloquent Peasant" is considered by some authorities as a prophecy, since the peasant is forced to deliver speeches, saying: "Not shall the one be silent whom thou hast forced to speak." This compulsion to speak in the name of the divine is called by some scholars the "prophetical condition."

In a Hittite text King Mursilis II (reigned *c.* 1334–*c.* 1306 BC) mentions the presence of prophets, but there is no information about the type of prophecy. More informative are texts from Mari (Tall al-Ḥarīrī, 18th century BC) in northwest Mesopotamia, where some striking parallels to Hebrew prophecy have been discovered. The Mari prophets—believed to be inspired—spoke the word of the god Dagon just as Israelite prophets spoke the word of Yahweh.

In Mari, the two key words for prophet are *muḫḫum* (an ecstatic, a frenzied one) and *āpilum* (the one who responds). Both may be connected with the cult, but there are incidents indicating that the *muḫḫum* was not bound to the cultic setting but received his message in a direct revelation from his god. The *āpilum* usually acted within a group of fellow prophets. Many of their sayings are political in nature, but there are also oracles that deal with the king's duty to protect the poor and needy, indicating that an ethical dimension was present among the Mari prophets. The messages could also contain admonitions, threats, reproofs, accusations, and predictions of either disaster or good fortune.

The Mari texts are important in the history of prophecy because they reveal that inspired prophecy in the ancient Middle East dates back 1,000 years before Amos and Hosea (8th century BC) in Israel. From Mesopotamia there is evidence of the *maḫḫu*, the frenzied one, known in Sumerian texts as the *lú-gub-ba*. Mention also is made of some prophets who spoke to Assyrian kings, and their message is sometimes introduced with the clause: "Do not fear." Omina (omens) texts containing promises or predictions are also known. In one of the *maqlu* ("oath") texts, in which an *āšipu* priest is being sent forth by his god, the deity first asks "Whom shall I send?"

The *baru* (a divinatory or astrological priest) declared the divine will through signs and omens, and thus by some is considered to have been a prophet. Though he might possibly have had visions, he was not in actuality an ecstatic. The art of divination became very elaborate in the course of time and required a long period of training.

Zoroaster, the 7th–6th-century-BC Iranian founder of the religion that bears his name, is one of the least well-known founders of a religion because of the character of the existing textual materials and because some scholars have advocated that Zoroaster is a mythical figure. He may have been, however, an ecstatic priest-singer, or *zaotar,* who used special techniques (especially intoxication) to achieve a trance. Zoroaster found the priests and cult of his day offensive, and opposed them. He preached the coming of the kingdom of the god Ahura Mazdā (Ormazd), who is claimed to have revealed to Zoroaster the sacred writings, the Avesta. In the *Yasna* (a section of the Avesta), Zoroaster refers to himself as a Saoshyans, a saviour. Messianic prophecies of the end of the world are found in Zoroastrian literature, but these are more a literary product than actual prophetic utterance (see also ZOROASTRIANISM AND PARSIISM).

Prophets were a common phenomenon in Syria-Palestine. In an Egyptian text (11th century BC), Wen-Amon (a temple official at Karnak) was sent by the pharaoh to Gebal (Byblos) to procure timber. While there, a young noble of that city was seized by his god and in frenzy gave a message to the king of Gebal that the request of Wen-Amon should be honoured. In another instance, an Aramaic inscription from Syria records that the god Ba'al-shemain told King Zakir (8th century BC) through seers and diviners that he would save the king from his enemies. These chapters reveal the close connection between sacrificial rites and divine inspiration. In the Old Testament book of Numbers, chapters 22–24, the Mesopotamian prophet Balaam (who may have been a *maḫḫu*) from Pethor, whom the Moabite king Balak had asked to curse the invading Israelites, is mentioned. In chapter 27, verse 9, of Jeremiah, another Old Testament book, it is said that prophets, diviners, and soothsayers were in the neighbouring countries of Judah: in Edom, Moab, Ammon, Tyre, and Sidon. Since so little is known about these prophets, the question of the uniqueness of Hebrew prophecy is difficult to assess (see also MIDDLE EASTERN RELIGIONS).

Origins and development of Hebrew prophecy. The Hebrew word for prophet is *navī',* usually considered to be a loan word from Akkadian *nabū, nabā'um,* "to proclaim, mention, call, summon." Also occurring in Hebrew are *ḥoze* and *ro'e,* both meaning "seer," and *nevi'a* ("prophetess").

Though the origins of Israelite prophecy have been much discussed, the textual evidence gives no information upon which to build a reconstruction. When the Israelites settled in Canaan, they became acquainted with Canaanite forms of prophecy. The structure of the prophetic and priestly function was very much the same in Israel and Canaan. Traditionally, the Israelite seer is considered to have originated in Israel's nomadic roots, and the *navī'* is considered to have originated in Canaan, though such judgments are virtually impossible to substantiate. In early Israelite history, the seer usually appears alone, but the *navī'* appears in the context of a prophetic circle. According to I Samuel, there was no difference between the two categories in that early time; the terms *navī'* and *ro'e* seem to be synonymous. In Amos, *ḥoze* and *navī'* are used for one and the same person. In Israel, prophets were connected with the sanctuaries. Among the Temple prophets officiating in liturgies were the Levitical guilds and singers: the "sons" of Asaph, Heman, Jeduthun, who are said to "prophesy with lyres, with harps, and with cymbals" (I Chronicles). Other prophetic guilds are also mentioned. Members of these guilds generally prophesied for money or gifts and were associated with such sanctuaries as Gibeah, Samaria, Bethel, Gilgal, Jericho, Jerusalem, and Ramah. Jeremiah mentions that the chief priest of Jerusalem was the supervisor of both priests and prophets, and that these prophets had rooms in the Temple buildings. In pre-Exilic Israel (before 587/586 BC), prophetic guilds were a social group as important as the priests. Isaiah includes the *navī'* and the *qosem* ("diviner," "soothsayer") among the leaders of Israelite society. Divination in the pre-Exilic period was not considered to be foreign to Israelite religion.

Prophets connected with sanctuaries

In reconstructing the history of Israelite prophecy, the prophets Samuel, Gad, Nathan, and Elijah (11th to 9th centuries BC) have been viewed as representing a transitional stage from the so-called vulgar prophetism to the literary prophetism, which some scholars believed represented a more ethical and therefore a "higher" form of prophecy. The literary prophets also have been viewed as being antagonistic toward the cultus. Modern scholars recognized, however, that such an analysis is an oversimplification of an intricate problem. It is impossible to prove that the *nevi'im* did not emphasize ethics simply because few of their utterances are recorded. What is more, none of the so-called "transitional" prophets was a reformer or was said to have inspired reforms. Samuel was not only a prophet but also a priest, seer, and ruler ("judge") who lived at a sanctuary that was the location of a prophetic

guild and furthermore was the leader of that *navi'* guild. In the cases of Nathan and Gad there are no indications that they represented some new development in prophecy. Nathan's association with the priest Zadok, however, has led some scholars to suspect that Nathan was a Jebusite (an inhabitant of the Canaanite city of Jebus).

Elijah was a "prophet father" (or prophet master) and a prophet priest. Much of his prophetic career was directed against the Tyrian Baal cult, which had become popular in the northern kingdom (Israel) during the reign (mid-9th century BC) of King Ahab and his Tyrian queen, Jezebel. Elijah's struggle against this cult indicated a religio-political awareness, on his part, of the danger to Yahweh worship in Israel; namely, that Baal of Tyre might replace Yahweh as the main god of Israel.

Classical prophecy in Israel and Judah

The emergence of classical prophecy in Israel (the northern kingdom) and Judah (the southern kingdom) begins with Amos and Hosea (8th century BC). What is new in classical prophecy is its hostile attitude toward Canaanite influences in religion and culture, combined with an old nationalistic conception of Yahweh and his people. The reaction of these classical prophets against Canaanite influences in the worship of Yahweh is a means by which scholars distinguish Israel's classical prophets from other prophetic movements of their time. Essentially, the classical prophets wanted a renovation of the Yahweh cult, freeing it from all taint of worship of Baal and Asherah (Baal's female counterpart). Though not all aspects of the Baal-Asherah cult were completely eradicated, ideas and rituals from that cult were rethought, evaluated, and purified according to those prophets' concept of true Yahwism. Included in such ideas was the view that Yahweh was a jealous God who, according to the theology of the psalms, was greater than any other god. Yahweh had chosen Israel to be his own people and, therefore, did not wish to share his people with any other god. When the prophets condemned cultic phenomena, such condemnation reflected a rejection of certain kinds of cult and sacrifice—namely, those sacrifices and festivals not exclusively directed to Yahweh but rather to other gods. The prophets likewise rejected liturgies incorrectly performed. The classical prophets did not reject all cults, per se; rather, they wanted a cultus ritually correct, dedicated solely to Yahweh, and productive of ethical conduct. Another important concept, accepted by the classical prophets, was that of Yahweh's choice of Zion (Jerusalem) as his cult site. Thus, every cult site of the northern kingdom of Israel and all the sanctuaries and *bamot* ("high places") were roundly condemned, whether in Israel or Judah.

Amos, whose oracles against the northern kingdom of Israel have been misunderstood as reflecting a negative attitude toward cultus per se, simply did not consider the royal cult of the northern kingdom at Bethel to be a legitimate Yahweh cult. Rather, like the prophet Hosea after him, Amos considered the Bethel cult to be Canaanite.

Prophets of the ancient Middle East generally interjected their opinions and advice into the political arena of their countries, but in this regard the classical Hebrew prophets were perhaps more advanced than other prophetic movements. They interpreted the will of God within the context of their particular interpretation of Israel's history, and on the basis of this interpretation often arrived at a word of judgment. Important to that interpretation of history was the view that Israel was an apostate people—having rejected a faith once confessed—from the very earliest times, and the view that Yahweh's acts on behalf of his chosen people had been answered by their worship of other gods. In this situation, the prophets preached doom and judgment, and even the complete destruction of Israel. The source of prophetic insight into these matters is the cultic background of liturgical judgment and salvation, wherein Yahweh judged and destroyed his enemies, and in so doing created the "ideal" future. What is totally unexpected is that the prophets would go so far as to include Israel itself as among Yahweh's enemies, thus using these ideas against their own people. Usually, however, the prophets allowed some basis for hope in that a remnant would be left. The future of this remnant (Israel) lay in the reign of an ideal king (as described in Isaiah), indicating that the prophets were not antiroyalists. Though they could and did oppose individual kings, the prophets could not make a separation between Yahweh and the reign of his chosen king or dynasty. Their messianic ideology, referring to the messiah, or anointed one, is based on old royal ideology, and the ideal king is not an eschatological figure (one who appears at the end of history). In this respect, the prophets were nationalistic; they believed that the ideal kingdom would be in the promised land, and its centre would be Jerusalem.

Concepts of the remnant and of the messiah

With the Exile of the Judaeans to Babylon of 586 BC, prophecy entered a new era. The prophecies of what is called Deutero-Isaiah (Isaiah 40–45), for instance, were aimed at preserving Yahwism in Babylonia. His vision of the future went beyond the pre-Exilic concept of a remnant and extended the concept into a paradisiacal future wherein Yahweh's new creation would be a new Israel. This tone of optimism is continued in the prophetic activity (late 6th century BC) of Haggai and Zechariah, prophets who announced that Yahweh would restore the kingdom and the messianic vision would come to pass. Prerequisite to this messianic age was the rebuilding of the Temple (which was viewed as heaven on earth). When, however, the Temple had been rebuilt and long years had passed with neither the kingdom being restored nor the messianic age initiated, Israelite prophecy declined.

There is a tendency in prophetic preaching to spiritualize those aspects of religion that remain unfulfilled; herein lie the roots of eschatology, which is concerned with the last times, and apocalyptic literature, which describes the intervention of God in history to the accompaniment of dramatic, cataclysmic events. Since the predictions of the classical prophets were not fulfilled in a messianic age within history, these visions were translated into a historical apocalypse, such as Daniel. Why prophecy died out in Israel is difficult to determine, but Zechariah offers as good an answer as any in saying that the prophets "in those days" told lies. Prophets did appear, but after Malachi none gained the status of the classical prophets. Another reason may be found in Ezra's reform of the cult in the 5th century BC, in which Yahwism was so firmly established that there was no longer any need for the old polemics against Canaanite religion.

Reinterpretation of the prophetic heritage

Prophecy and apocalyptic literature. With the advent of post-Exilic Judaism (Ezra and after), including its emphasis on law and cult, there was not much room left for prophecy. The prophetic heritage was channelled through the teaching of their words. What remained of prophetic activity was expressed in various literary works that claimed esoteric knowledge of the divine purpose. The apocalyptic writers saw themselves as taking over and carrying on the prophetic task, but they went beyond the prophets in their use of old mythological motifs. The events they described had usually occurred long ago, but their recounting of these events was for the purpose of hinting and even predicting the events of the future. There was a far greater emphasis upon predictive speculation about the future than on the prophetic analysis and insight into history. The apocalyptic authors wrote pseudonymously, using the names of ancient worthies (such as Adam, Enoch, Abraham, Daniel, and Ezra). The literature is predominantly prose, but that of the classical prophets was predominantly poetry. Apocalyptic language is lavish in its use of fantastic imagery, frequently using riddles and numerical speculations. In apocalyptic literature angelology came into full blossom, with accounts of fallen angels (fallen stars) caught up in the forces opposed to God, frequently pictured in the old mythological motif of the struggle between darkness and light. Wild beasts symbolized peoples and nations, and there were esoteric calculations and speculations about the different eras through which history was passing as the world approached the eschaton (the consummation of history).

Dominant in apocalyptic literature is the theme of God's sovereignty and ultimate rule over all the universe. The message of the apocalyptic writers is one of both warning of the doom to come at the end of history, and hope in the new age beyond history under the rule of God, when the righteous will be vindicated (see above, *Eschatology*).

Prophecy and prophetic religion in postbiblical Judaism. Though prophecy did not cease functioning in early Judaism, rabbinical Judaism—that influenced by rabbis, scholars, and commentators of the Bible—sought to limit it by advocating the pre-Exilic era as the classical time of prophecy. Prophecy was not suppressed, but it came to be encircled by the law (Torah) in that all prophecy had to be in harmony with Torah, which was the definitive revelation of God's will. Thus, rabbinical Judaism gave prophecy its place of importance, but only as a phenomenon of the past. Such a theological stricture could not restrain the charismatic, eschatologically oriented patriots who arose during the time of Roman hegemony (mid-1st century BC–4th century AD). One rabbi, Akiba ben Joseph, joined with a messianic pretender, Bar Kokhba (originally Simeon ben Koziba) in a revolt (132–135) and functioned as a prophet within that movement.

Eschatologically oriented prophets

Some prophets are known from the period of Hellenistic Judaism. I Maccabees, chapter 14, relates that Simon Maccabeus, who finally secured political independence for Judaea in 141/140 BC, was chosen as "leader and high priest forever, until a trustworthy prophet should arise." The same notion of a prophet soon to appear is expressed in chapter 1 of I Maccabees. The Hasmonean (Maccabean) prince John Hyrcanus (reigned 135/134–104 BC) was regarded as fulfilling these expectations and was called a prophet by the 1st-century AD Jewish historian Josephus (*Jewish War*). Josephus also mentions some Zealots (Jewish revolutionaries) as prophets and also one Jesus, son of Ananias, who in AD 62 predicted the destruction of the Temple and the defeat of the Jews. Josephus also mentions the seer Simon, a prophet leader (*Antiquities*), and Menahem, who prophesied in the 1st century BC. Among the followers of Judas Maccabeus, the leader of the 2nd-century-BC revolt, there apparently were persons who divined knowledge of the future. These and other notations indicate that seers and prophets played an important role in the intertestamental and postbiblical periods.

Jewish theology in Alexandria (Egypt) took up early rabbinical ideas and postulated that the will of God was to be discerned in the Torah and affirmed that the interpretation of law succeeded both the prophetic office and the role of sages. The law was thus considered to be superior to prophetic teaching. The Jewish philosopher Philo of Alexandria (*c.* 30 BC–after AD 40) affirmed that the Jews are a people of prophets. He also asserted that when a prophet has reached the fourth and final stage of ecstasy he is ready to become an instrument of divine power. Though Philo was influenced by Hellenistic concepts of prophecy, his basic foundation was still the Old Testament. Later rabbis believed that prophecy, though it was a gift from the world beyond, still required some knowledge. In rabbinic discussions of the nature of truth, it was generally held that reason alone was necessary but insufficient; prophecy could supply what was missing.

Superiority of law in relation to prophecy

The medieval Jewish philosopher Maimonides understood prophecy as an emanation from God to the intellect of man. Thus, prophecy could not be acquired by human effort. The divine gift of prophecy was bestowed upon those with both mental and moral perfection, combined with the presence of superior imagination. Opponents of this view advocated that Maimonides' concept of prophecy was not Jewish because Jewish prophecy always showed itself to be miraculous (see also JUDAISM).

PROPHECY IN CHRISTIANITY

Divination and prophecy in the Hellenistic world. The problem of false prophets that occurred in Old Testament times also occurred in the early Christian communities. Prophets and diviners were widespread throughout the Hellenistic world. The Greek *prophētēs* was not only a forthteller but also an interpreter of divine messages. In addition, there were mantics (from the Greek *mantis*)—*i.e.,* visionary seers—whose visions were interpreted by prophets, soothsayers, diviners of all kinds, and especially astrologers. The impetus for much of this activity came from Babylonia. The influx of new religions from the East brought a profusion of astrologers and prophets. Many schools of astrology were founded throughout the Hellenistic world, and old schools of philosophy became very much occupied with astrology.

New Testament and early Christianity. Prophecy in the New Testament is seen as both a continuation of Old Testament prophecy as well as its fulfillment. For New Testament authors, the correct interpretation of Old Testament prophecy is that it speaks *in toto* of Christ. To prove their point, they often cite passages of Old Testament prophecy that are then elucidated as the words of God about Christ. New Testament writers follow Jesus himself in this matter, and Jesus is taken to be the prophet that was promised in Deuteronomy (see John 1:45, *cf.* 5:39, 6:14; Acts 3:22 *ff.*). Jesus regarded himself as a prophet, and so did some of his contemporaries. One special aspect of the prophetic image, however, is missing in Jesus: he was not an ecstatic, although supernatural revelations are found in connection with him; *e.g.,* the transfiguration of Jesus as witnessed by some of his Apostles on Mt. Tabor. In these New Testament descriptions of the transfiguration, Jesus is proclaimed to be the Son of God in words borrowed directly from Old Testament enthronement ritual. As a prophet, Jesus predicted his own death, his return as the Son of man at the end of the world, and the destruction of the Jerusalem Temple. At many points, Jesus is compared with and interpreted by the classical prophets in New Testament writings: his death—seen as the martyrdom of a prophet, his sufferings, and even his identity.

Though the New Testament describes Jesus as a prophet, he is at the same time believed to be more than a prophet: he is the expected Messiah (Greek *christos,* "anointed one"), predicted by prophets of old, who should reign as the Son of David and the Son of God. The royal ideology of the Old Testament was most important to early Christianity, for herein lay the seeds of its doctrines of Christ (see above, *Eschatology*).

Several prophets are mentioned in the New Testament. One, Zechariah, is said to have perished "between the altar and the sanctuary" (Luke). Reference to his death is included by the Gospel writers because he was the last prophet before Jesus to have been killed by the Jews. Zechariah, the father of John the Baptist, uttered the Benedictus ("Blessed," the initial Latin word of the prophetic song) under the inspiration of the spirit. His wife, Elizabeth, also was described as being inspired by the spirit.

Others are Simeon, the prophetess Anna, and John the Baptist. These prophets are conceived by the New Testament writers as the termination of Old Testament prophecy, a concept also expressed by Jesus with reference to John the Baptist.

New Testament prophetic figures

The New Testament mentions several prophetic figures in the early church. Among them are Agabus of Jerusalem; Judas Barsabbas and Silas, who also were elders of the Jerusalem Church; the four prophesying daughters of Philip the evangelist; and John, the author of Revelation. The term prophet is used with reference to an office in the early church along with evangelists and teachers, and the recipient of the letter bearing his name, Timothy, is called both a minister and a prophet. The prophet's role in the early church was to reveal divine mysteries and God's plan of salvation. Paul instructed his followers in the correct use of prophecy, and evaluated it as more beneficial to the life of congregations than ecstatic glossolalia (speaking in tongues). He considered prophecy to be the greatest spiritual gift from God, and in his view a prophet therefore ranks ahead of evangelists and teachers. With all this prophetic activity, the problem of false prophecy was crucial, and warnings against it abound in the New Testament. The most dangerous of the false prophets is predicted in the book of Revelation to John as yet to come. Many of these prophets, viewed as magicians and exorcists, are condemned for inducing chaos and for leading people astray. Therefore all prophetic activity had to be examined.

In the period immediately after the Apostles, prophets continued to play an important leadership role in the church, sometimes being called high priests. They were the only ones permitted to speak freely in the liturgy because of their inspiration by the Holy Spirit. Gradually, however, the liturgy became more and more fixed, and

less freedom and innovation was permitted; this change, combined with the threat of false prophecy, eliminated these charismatic personalities. Among the heretical sects
Montanism that advocated a return to prophetic activity, Montanism (2nd century), led by the prophet Montanus, advocated that the spirit of truth had come through Montanus. The freedom of doctrinal innovation that Montanus advocated could well have led to doctrinal anarchy, and the result of the struggle against this heresy was the suppression of charismatic prophecy, wherein ecstatic inspiration came to be viewed by the church as demonic.

Another prophet who created a problem in the early church was Mani—the 3rd-century founder of a dualistic religion that was to bear his name (Manichaeism)—who considered himself to be the final messenger of God, after whom there was to be no other.

Prophetic and millenarian movements in later Christianity. In Western medieval church doctrines and rituals, active prophecy had no place. Prophetic activity was carried on, however, through holy orders. Mystically oriented holy men would sometimes appear as prophets with a special message, and even ecstatics found their places within the monasteries. In Eastern Christianity, monastic life stressed training in mystical experience.

Throughout Christian history there have been millenarian movements, usually led by prophetic-type personalities and based on the New Testament belief in Christ's return. Their basic doctrine is chiliasm (from Greek *chilioi,* "thousand"), which affirms that Christ will come to earth in a visible form and set up a theocratic kingdom over all the world and thus usher in the millennium, or the 1,000-year reign of Christ and his elect.

The early and medieval church hierarchy generally opposed chiliasm because such movements often became associated with nationalistic aspirations. Though the key leaders of the Protestant Reformation opposed chiliasm, and therefore minimized its effects upon the emergent denominations (*e.g.,* Lutheran, Calvinist, and Anglican), chiliasm did influence Anabaptist circles (radical reformation groups), and through them chiliastic ideas influenced Protestant Reformed theology and have appeared in reform movements, such as Pietism in Lutheran churches, and various revivalistic movements.

PROPHECY IN ISLĀM

The centrality of prophecy in Islām. Pre-Islāmic prophecy in Arabia was no different in character from other Semitic prophecy. Pre-Islāmic terms for prophet are *'arrāf* and *kāhin* ("seer," cognate to Hebrew *kohen,* "priest"). The *kāhin* could often be a priest, and as a
The role of the *kāhin* diviner he was an ecstatic. The *kāhin* was considered to be possessed by a *jinnī* ("spirit"), by means of whose power miracles could be performed. Also, poets were considered to be possessed by a *jinnī* through whose inspiration they composed their verses. The importance of the seers and diviners was noted in all aspects of life. Any problem might be submitted to such men, and their oracular answers were given with divine authority. It is not surprising, therefore, to find that a *kāhin* often became a sheikh, a temporal leader, and there were instances in which the position of *kāhin* was hereditary.

It was against this background that the founder of Islām, Muḥammad, appeared. During his early career in Mecca (in Arabia) he was considered by his tribesmen, the Quraysh, to be only another *jinnī*-possessed *kāhin.* His utterances during this time were delivered in the same rhymed style as that used by other Arab prophets and were mostly the products of ecstatic trances. At about 40 years of age Muḥammad experienced the promptings of the one god, Allāh, and retreated into the solitude of the mountains. These retreats served psychologically as preparations for his later revelations. The central religious problem of Muḥammad was the fact that Jews had their sacred scriptures in Hebrew, and Christians had theirs in Greek, but there was no written divine knowledge in Arabic. Muḥammad's preoccupation with this concern, along with a sense of the coming Day of Judgment, became the seeds of his new religion. Contemplation had matured Muḥammad, and biographers point out that, as one may conclude from the Qur'ān, Muḥammad received the divine call in a vision. His ecstatic revelations were in the form of auditions, usually involving the angel Gabriel reading the divine message from a book. The illiterate Muḥammad had his wife Khadījah, who was 15 years his senior, record them, and they are preserved in the Qur'ān. Because this is believed to be a verbatim copy of the Heavenly Book, literally the words of Allāh himself, it cannot be questioned.

Muḥammad considered himself to be more than a mere prophet (*nābi*); he thought of himself as the messenger (*rasūl*) of Allāh, the final messenger in a long chain that had begun with Noah and run through Jesus. As Allāh's *rasūl,* Muḥammad saw his first mission to be that of warning the Arab peoples of the impending doomsday. No doubt Muḥammad was influenced by the Judeo-Christian tradition in his concept of the Day of Judgment, as well as in his concept of himself as a prophet. Muḥammad, who had felt at one time that Arabs were religiously inferior to Jews and Christians, became the medium of revelations that created Islām and raised the Arabs in Muḥammad's own evaluation to a status equal with that of the other two religions.

After AD 622, when Muḥammad left Mecca and found refuge in Medina, ecstatic revelations began to play a secondary role in his prophecy—due to his political concerns—and not only does the rhymed prose of his message give way to more conventional prose but the content is more obviously the product of reasoned reflection on all aspects of life.

The Qur'ānic doctrines of prophecy. An official Islāmic view, and also that of Muḥammad himself, was that Muḥammad was the final Prophet. The Qur'ān mentions
The final Prophet those men who are considered to have imparted divine knowledge: Adam, Noah, Abraham, Isaac, Jacob, Moses, David, and Jesus. None of these revealed Allāh's message in full, since they were sent only to one nation. Muḥammad, on the other hand, was sent to all nations and also to the *jinn.* The messages of the prophets before Muḥammad were believed to have been either forgotten or distorted, but Islām claims that the Qur'ān both corrects and confirms the sayings of the earlier prophets; Muḥammad is the "seal of the prophets"; *i.e.,* the end of prophecy. All prophecy before Muḥammad is incomplete and points to the coming of the final revelation.

The prophetic activity of Muḥammad serves as the foundation of Islām and Muslim society. The incomparable revelations of Muḥammad are believed to have brought true monotheism into the world, to which nothing can be added or taken away. Thus, there is no more need of prophets or revelations.

Later theological and philosophical doctrines. After the death of Muḥammad, the expansion of Islām brought it into contact with the world at large, and a Muslim culture (involving science, philosophy, and literature) emerged, partially as a result of the Muslim acquisition of Byzantine culture. Christians and Jews became advisers and officials in Muslim courts. Christian philosophers introduced Muslim students to the works of the 4th-century-BC Greek philosopher Aristotle and to Neoplatonism (a philosophical system concerning the complex levels of reality), to theories about the nature of man, to theology, to the nature of existence, and to cosmology. Philosophical discussions about God, however, leave little or no room
Displacement of prophets for prophets, and the savant displaced the prophet as the one proclaiming the will of God. As religious leaders, the savants were the keepers of *sunnah* (the life and habits of the prophet) and *ḥadīth* (traditions about Muḥammad's utterances and actions), which are supplements to the Qur'ān. Study of *ḥadīth* and *sunnah* contributed to the beginning of scholarly and scholastic activities in Islām, from which study emerged the Muslim system of duties and obligations (*fiqh*). Muslim theology began in the formulation of the doctrine of the general consensus (*ijmā'*), which was used to determine what was genuine *sunnah.* None ventured to question that Allāh was the only God, that Muḥammad was his prophetic messenger, or that the Qur'ān was Allāh's word; to have done so would have been tantamount to admitting that one was not a Muslim.

Scholastic philosophy was first introduced openly into Muslim theology by al-Ash'ari (10th century) who was the first to give Islām a systematic exposition. Another theologian, Ibn Sīnā (Avicenna), considered prophecy still to be a fundamental aspect of Islām, but for him, a prophet was not the spirit-possessed spokesman of God but rather an intelligent, intuitive man whose insight results in a place of leadership in society. Another philosopher, Ibn Rushd (Averroës), denied the belief that man's knowledge could ever be the same as God's knowledge; he also denied doctrines of predestination and corporeal resurrection, both of which were aspects of Muḥammad's message.

Prophetic figures after Muḥammad. The fact that Muḥammad was considered to be the final prophet did not end prophecy in Islām. After Muḥammad's death, several seers proclaimed themselves his successors. Muḥammad had designated no one to succeed himself, and left no sons. Abū Bakr, the father of Muḥammad's wife 'A'ishah, was chosen caliph (Arabic *khalīfah*, "substitute, deputy"), but this did not discourage others from claiming that they were called of Allāh and thus trying to lead their own tribes as Muḥammad had led his. Such movements were crushed by force, which contributed to the rapid expansion of Islām.

The *mahdī*

Some prophets claimed that they were long-awaited saviour-deliverers (*mahdī*, "restorer of the faith") and even gained some following beyond their own local tribes. Muḥammad Aḥmad ibn as-Sayyid 'Abd Allāh of the Sudan preached a holy war against Egypt (1881) and fought and defeated the British governor-general C.G. Gordon at Khartoum in 1885. In India (Punjab), Mirza Ghulan Aḥmad claimed that he had received the spirit of Jesus and that he was a prophet-messiah. He recorded his revelations from Allāh in a book. Considering himself to be the Christ to his generation, he set out to reform Islām by liberalizing strict orthodoxy, yet avoiding the extremes of the pro-Western movements of his time. He gained a large following among middle-class Muslims, but was soon disowned by orthodox Islām. His sect (Ahmadiyah), though small in numbers, has through its missionary activities spread over much of the world. Its sociopolitical stance is similar to that of the Black Muslims of the United States (see also ISLĀM).

PROPHECY IN OTHER RELIGIONS

Prophetic movements and figures in the Eastern religions. Buddhist literature contains predictions of a certain Buddha Maitreya, who will come as a kind of saviour-messiah to inaugurate a paradisaical age on earth. Gautama the Buddha himself, the 6th-century-BC founder of Buddhism, mentioned this prediction. Among the Hindus, the *Purāṇa* literature ("old history") contains prophetic passages, but these are to be understood as predictions after the event has occurred. Hindu religion has had many prophetic reformers, and the tribes of India, in their struggle for freedom, have produced prophets who combined the ideas of religious freedom with the hope of political and social freedom. The Oraons, a tribe in Chota Nāgpur, saw several prophets (*bhagats*) appear around the turn of the 20th century. Their intent was to free their people from foreign culture and political rule, returning to the older Hindu culture and religion. Such efforts often led to armed rebellion and ended in disaster.

In ancient China, divination was commonplace. One Confucian book involving divination, the "Classic of Changes," may have been connected with pre-Han Confucianism (before the 3rd century BC). Classical Confucian religion, however, emphasized the importance of rational process over inspiration and divination. Autocratic governments eliminated any such revolutionary, prophet-led movements as occurred in India, and any prophecy against the establishment was regarded as heretical. Inspired prophecy found little place in the official state religion. This situation did not rule out prophecy in folk religion, in which prophets appeared and promised their followers the good life in this world and in the next. In modern times, some of these movements became religio-political movements, as when Hung Hsiu-Ch'üan, an ecstatic epileptic noble of the middle 19th century, started a movement called the Taiping ("Great Peace"), a sect claiming that it was establishing the correct political order anew. Hung's movement—perhaps under the impact of Protestant missions—was quite austere, and it opposed magic, idols, and belief in spirits. He considered the New Testament to be authoritative for his new sect, and its rapid growth—aided by connections with other revolutionary movements—soon resulted in a genuine danger to the Manchu ruler of China. The Taiping Rebellion was crushed by Gordon in 1864.

Folk religion prophets

Diviners and shamans (male and female) are well represented in old Japanese Shintō. Japanese shamanism, which was closely related to Korean shamanism, often played a role in political disturbances and still does. Among old Japanese Buddhist sects is that founded by Nichiren (13th century AD), a prophetic enthusiast, religious revivalist, and zealous nationalist who taught that the Japanese people were the chosen people of God. In the Shintō revival movements of the 18th and 19th centuries, inspired persons with eschatological concepts founded movements that became messianic in character, and drew many of their followers from among the farmers, many of whom had practiced a Buddhist folk piety.

Prophetic movements and figures in the religions of nonliterate cultures. In many nonliterate cultures, especially those of Africa, shamans, seers, and prophets are quite common. The same distinction between technical divination and charismatic prophecy is to be found in these cultures as in the ancient Middle East. When it is possible to trace the history of prophetic activity in Africa, scholars usually find that it arises in times of confrontations with foreign cultures and with the advent of new religions. A sharp distinction between the diviner and the prophet cannot always be maintained, for diviners sometimes appear as prophets. A diviner may hear the voice of a god or spirit in his dreams and visions (in Zulu he is called a "dreamhouse") and receive a message. Some prophets, avowing a call, deny any training in prophecy. There are many parallels with the "rebel" prophets of India. Ecstatic prophets have played an important role not only in chiliastic and messianic movements but also in those movements opposing imperialism and European colonization of Africa. Their goal was and is a return to the old African culture and religion. Eschatological motifs have often been used in the prophetic preaching of tribal and national movements aspiring for freedom. Many of these prophets took up Christian ideas. Nxele, a 19th-century prophet of the South African Xhosas, preached the return of the dead on a certain day, and his successor, Mlandsheni, claimed to be the reincarnation of Nxele. He and others like him were healers and miracle workers.

Some of the prophetic founders of reform movements, which often were more political than religious, became messianic figures. Other prophets started out as Christian converts but came to a strong awareness that God had destined them to separate from their churches and lead syncretistic movements (fusions of various sources), all of which incorporate aspects of old African religion and, often, allow polygamy. In all these movements, syncretistic or not, there are also many prophetesses.

Prophets also have been found among American Indians. In 1675 a medicine man, Popé, arose as a prophetic leader among the Pueblo Indians. He preached the end of Spanish tyranny and a restoration of Indian sovereignty. At the height of the movement, several massacres took place, along with the burning of various church buildings.

(G.W.A.)

Miracle

Miracles are extraordinary and astonishing happenings that are attributed to the presence and action of an ultimate or divine power. This section treats the nature, functions, and sources of miracles; it then attempts to discover the place of miracles in the religions of the world and the variety of interpretations given to them. It considers miracles mainly as seen, experienced, and interpreted in the religious context, irrespective of the scientific or philosophical judgment passed on them from the outside.

NATURE AND SIGNIFICANCE

Etymology and definition

A miracle is generally defined, according to the etymology of the word—it comes from the Greek *thaumasion* and the Latin *miraculum*—as that which causes wonder and astonishment, being extraordinary in itself and amazing or inexplicable by normal standards. Because that which is normal and usual is also considered as natural, miracles have occasionally been defined as supernatural events, but this definition presupposes a very specific conception of nature and natural laws and cannot, therefore, be generally applied. The significance of a miraculous event is frequently held to reside not in the event as such but in the reality to which it points (*e.g.*, the presence or activity of a divine power); thus, a miracle is also called a sign—from the Greek *sēmeion* (biblical Hebrew *ot*)—signifying and indicating something beyond itself. Extraordinary and astonishing occurrences become specifically religious phenomena when they express, reveal, or signify a religious reality, however defined.

Belief in miraculous happenings is a feature of practically all religions, and the incidence of miracles (*i.e.*, of belief in and reports regarding miracles) is universal, though their functions, nature, purpose, and explanations vary with the social and cultural—including theological and philosophical—context in which they appear. However inexplicable, all miracles have an explanation in the sense that they are accounted for in terms of the religious and cultural system that supports them and that, in turn, they are meant to support. Without such an accompanying—explicit or implicit—theory (*e.g.*, the presence, activity, and intervention of such realities as gods, spirits, or magical powers), there would be no miracles in the aforementioned sense but only unexplained phenomena.

TYPES AND FUNCTIONS OF MIRACLES

There is no general rule determining the types of occurrences that can be classified as miracles; they vary according to the cultural matrix of beliefs and assumptions. The mythological accounts of the origins of the gods and their activities in the primeval past, as well as accounts of the activities of other primeval beings, such as first ancestors and culture heroes, should, perhaps, not be classed as miracles, and the term is better reserved for outer, objective events—as distinct from such phenomena as inner experiences and visions—that can be regarded as divine interventions or as manifestations of divine or supernatural powers. In many cultures, primitive as well as some that were more highly developed, such as the ancient classical and Oriental civilizations, the operation of extraordinary forces was taken for granted and was integrated into the total world picture and into the procedures and the modes of action—*e.g.*, magic, oracles, divination, and shamanism—of ordinary life. There were certain kinds of divine or spirit action and of cosmic operation that were considered to be a part of the normal order of things, even though it was generally admitted that priests and shamans would frequently resort to deception in their diverse activities, which included such manifestations as prophecy, oracles, healing, magic, and judgment by ordeal.

Purposes of miracles

Revelation and signification. The purpose of a miracle may be in the direct and immediate result of the event—*e.g.*, deliverance from imminent danger (thus, the passage of the children of Israel through the Red Sea in the Old Testament book of Exodus, chapter 14), cure of illness, or provision of plenty to the needy. Nevertheless, the ultimate purpose frequently is the demonstration of the power of the god or of the saint, the "man of God" through whom the god works, to whom the miracle is attributed. Thus, the crossing of the Red Sea by the Israelites is described not solely in terms of salvation from great danger but as a revelation of the saving presence of God and of the consequent obligation to serve and obey him; according to the account in Exodus: "and Israel saw the great work which the Lord did against the Egyptians, and the people feared the Lord; and they believed in the Lord and in his servant Moses." The purpose of a miraculous occurrence is thus often to reveal a divine reality or numinous dimension. The occurrence may be an event concerned with natural needs or situations, such as illness, hunger, or distress, or a specifically religious event that effects some form of salvation or revelation, such as the theophany on Mt. Sinai in which God gave to Moses the Ten Commandments, the Resurrection of Jesus Christ, or the revelation of the Qur'ān to Muḥammad. Even in these specifically religious events, the miraculous element is not necessarily of the essence but occurs as merely an accompanying circumstance designed to arrest the attention and to impress on everyone the unique character and significance of the occasion. Thus, theoretically at least, the theophany at Mt. Sinai could have taken place without thunder and lightning; Jesus need not have been born of a virgin; Muḥammad need not have made his miraculous journey to heaven. In actual fact, however, the very nature and quality of a religious event attracts miraculous elements, elaborations, and embellishments; and thus, for example, the founders of almost all religions are at the centre of great miracle cycles, and miracles occur as a rule in connection with persons and objects of religious significance, such as saints, sacraments, relics, holy images, and the like.

Authentication. In practice it is difficult to distinguish the revelatory or signifying miracles from miracles of authentication—*i.e.*, miraculous happenings that serve: (1) as credentials for claimants to religious authority in the form of leadership (*e.g.*, in Exodus, chapter 4, Moses convinces the Israelites of the authenticity of his mission by miraculous performances) or prophecy (*e.g.*, in Deuteronomy, chapter 18, it is said that a prophet is disqualified if the sign that he has predicted does not come to pass); (2) as the demonstration of the superior power of a particular god (*e.g.*, in Exodus, chapter 7, Aaron's staff swallowed up the staffs of the Egyptian magicians, which showed the superiority of the God of the Israelites); (3) as proof of the sanctity of a holy man, a holy site, or a holy object; or (4) more generally as evidence of the truth of a particular religion.

SOURCES OF MIRACLES

Spiritual sources. The source of miracles is always a divine, spiritual, supernatural, sacred, or numinous power that may be conceived in personal form (*e.g.*, God, gods, spirits) or impersonal form (*e.g.*, mana or magic). The sacred may manifest itself in natural phenomena, such as thunderstorms or earthquakes, that evoke appropriate feelings of awe, but these are not usually considered miracles unless attended by special circumstances—*e.g.*, being predicted by a "man of God" or coinciding with an event of religious significance. As reported in the Gospel According to Matthew, chapter 27, at the moment of Jesus' death on the cross, "the curtain of the temple was torn in two, from top to bottom; and the earth shook, and the rocks were split; the tombs also were opened, and many bodies of the saints who had fallen asleep were raised, and coming out of the tombs after his resurrection they went into the holy city and appeared to many." The belief that thunder and lightning are manifestations of divine powers is very common, and many deities have been interpreted as personifying them or at least as being symbolized by them. Even in the Old Testament, thunderstorms and lightnings appear as manifestations or messengers of God. In this respect, the account of the theophany granted to the prophet Elijah in I Kings, chapter 19, marks a milestone in the history of religions, for "behold, the Lord passed by, and a great and strong wind rent the mountains, and broke in pieces the rocks before the Lord, but the Lord was not in the wind; and after the wind an earthquake, but the Lord was not in the earthquake; and after the earthquake a fire, but the Lord was not in the fire; and after the fire a still small voice" in which Elijah heard God.

Natural phenomena accompanied by special circumstances

In most cases theophanies and divine manifestations occur for a specific purpose: giving laws (*e.g.*, Moses and the theophany at Mt. Sinai; events in the lives of Numa Pompilius of Rome, Minos of Crete, and Lycurgus of Sparta, the ancient lawgivers in classical legend); saving interventions (*e.g.*, the voices resounding from the temple of Athena Pronaea in Delphi that caused the Persians to retreat); and the founding of cults (*e.g.*, the appearances of Mary, the mother of Jesus, at Lourdes, France, and

Fatima, Portugal). Gods would appear to their devotees in visions and dreams, but these experiences should, perhaps, not be treated under the same general heading with other miracles. Immediate divine action was often perceived in omens preceding important undertakings, in apparently natural phenomena occurring providentially at critical moments or in miraculous—*i.e.,* sudden and seemingly impossible—cures. In most cases, however, such divine interventions took place through some form of mediation, human or inanimate.

Human and inanimate sources. Man can be the object of miracles, as when his disease is miraculously healed, or their subject, as when he performs miracles, such as healing others, in the name of whatever power is moving him. The two aspects cannot always be strictly distinguished, as is seen in the case of saints whose bodies are immune from corruption after death or founders of religions whose birth is attended by supernatural manifestations. Generally speaking, however, it is the role of holy personages—and of their tombs and relics—as sources of miracles that are of importance in the history of religions and more especially in the history of popular cults.

Subjects of legends

Founders of religions. The attitudes of the founders of the great religions toward miracles vary considerably, but all have become the subject of legends of the most fantastic kind in popular belief, and much of this legendary material has been subsequently canonized in scripture and tradition.

Holy persons. Much closer to the lives and devotion of ordinary folk than the superhuman figures of the founders are the saints, monks, ascetics, and diverse kinds of holy men and women. The attitude toward saints and their miracles is very much the same on the popular levels of all religions, although the theoretical interpretations on the more theological level vary considerably. In Far Eastern religions it is often difficult to distinguish between saints and hero gods, because great men of renowned virtue can be deified and venerated and even receive officially approved state cults. Miracles occur as a matter of course at their tombs and relics. In Muslim as well as in Christian belief, the occurrence of miracles is part of the requirements for official recognition of sainthood and is interpreted as a special intervention by God, who thereby manifests his esteem for the saint or, more essentially, his salvific presence as realized concretely in the life and virtues of the saint. In Indian—Hindu and Buddhist—belief, miraculous powers are the "natural" result of ascetic practice and spiritual realization and can thus be considered as an almost natural manifestation of magical or spiritual causes.

Relics, images, and amulets

Sacred objects. Because the life span even of saints is limited, most of the miracles attributed to them occur through their inanimate remains at their tombs or through their relics. These relics may be parts of their bodies—often deliberately dismembered for wider distribution, so that a bone may be in one place, a hair in another, and the heart someplace else—or objects or parts of objects associated with their lives (*e.g.,* the shroud of Christ or fragments of the True Cross).

Not all miracle-working objects of veneration are relics. Statues and icons can work miracles, and in many Christian countries images and icons of the Virgin Mary are especially famed for their miraculous virtues. In the Christian Middle Ages the veneration of the sacrament of the Eucharist brought about a proliferation of miracles. Here, as in the case of images, a distinction can be made between the magical character of folk beliefs and the diverse theological doctrines concerning these religious objects; only rarely have religious authorities opposed the cult of saints, images, and relics and the concomitant belief in miracles—an exception is classical Protestantism, which in the 16th century rejected such cults.

Although they are not strictly sources of miracles, talismans and amulets—*i.e.,* objects believed to possess magical virtues such as good luck or protection of the bearer or owner from all kinds of danger—should be mentioned in this connection. They are found in diverse forms and sizes and in all kinds of material.

Sacred places. Miracles are often connected with special sacred places. Normally these are natural shrines, such as sacred groves, or temples and sanctuaries in which a god or spirit lives or has manifested himself or in which his statue, symbol, holy objects, or relics are enshrined. Holy places, such as Mecca and the Ka'bah in Islām or the Buddhist stupas, are centres of pilgrimages and veneration because of their religious significance and the religious values that they symbolize and not necessarily because miracles are wrought there; yet, popular devotion associates miracles with many of these holy sites.

MIRACLES IN THE RELIGIONS OF THE WORLD

It has already been suggested that the mythologies of primitive and ancient religions should not be designated miraculous insofar as they deal with mythical origins and ages; frequently they attempt to explain how certain regularities and what is now considered the normal course of things have come into being. The crucial distinction lies between religion on the popular primitive level and the more highly developed forms of religious belief. The tendency of the former is to relate to a concrete, magical presence of the sacred and to envisage the possibility of using this presence for the achievement of such desired ends as healing, blessing, or success in an undertaking. The higher forms of religion—though recognizing miracles or even demanding dogmatic affirmation of belief in them—exhibit a far more differentiated and complex attitude.

Hellenistic religion

Hellenistic religion presents one of the best examples of a civilization in which miracles play a major part. The intervention of the gods in the affairs of the Homeric heroes takes place in a cosmos in which the divine and human spheres still interact. Later Hellenistic syncretism conceived of the sublunar world as a distinct sphere, though higher powers could miraculously irrupt into it. Miraculous cures (*e.g.,* at the sanctuary of Asclepius at Epidaurus), divine manifestations of various kinds (*e.g.,* voices, dreams, and theophanies), and even virgin births and resurrections were widely reported.

Religions of the East. In the great religions of the East the belief in miracles is closely connected with the theory that ascetic practices and the knowledge of mystical formulas, such as the Sanskrit *mantras,* can give the practitioner unlimited magical powers.

Religions of India. India has become the classic land of wonders not because of the accounts of fantastic actions of divine beings or semidivine heroes and avatars (incarnations of Hindu gods) related in Indian mythology but because both popular religion and philosophical theory set no bounds to the magical powers that can be attained by great ascetics and yogis (adherents of Yoga, the Hindu philosophy teaching the suppression of all activity of mind, body, and will in order that the self may realize its distinction from them and attain liberation). Even if these magical powers are considered insignificant in higher religion and spiritually negligible, their reality is never doubted. The *Upaniṣad* and the *Brāhmaṇa*—ancient Sanskrit writings of the Vedic period—may consider the heights of religious insight and mystical experience as man's supreme aim, but neither the later classical sources nor contemporary Hindu belief ever question the miraculous powers of a holy man. The same attitude is shared by the other religions of Indian origin: Jainism and Buddhism.

The Buddha

The Buddha himself refused to spread his teaching by impressing his audience with miracles. According to the Aṅguttara Nikāya, one of the collections of the Buddha's sayings, there are three kinds of miracles—the miracle of magic, the miracle of thought reading, and the miracle of instruction—and of these the last is the most wonderful and excellent, whereas the other two are not much better than a conjuror's tricks. Yet the same text also describes what is implied by the miracle of magic: "there is one who, . . . having been one becomes many, . . . appears and vanishes, unhindered he goes through walls. . . . He dives in and out of the earth as if it were water. Without sinking he walks on water as if on earth. Seated cross-legged he travels through the sky like a winged bird. With his hand he touches and strokes the sun and the moon. . . ." The same text also asserts that not only was Gautama endowed

with these powers but so also were hundreds of monks of his order.

Religions of China. In China, although Confucianism in the strict sense has little room for miraculous elements, Taoism has produced a rich crop of thaumaturgy and magic on all levels of folk religion. No doubt the teaching of the Tao (literally, the Way) can be interpreted in terms of a sublime moral and perhaps even mystical doctrine. In actual fact it was one of the main sources of Chinese magic in all its forms, including the quest for the elixir of life. Religious Taoism, with its theory of a balance and interaction of cosmic forces, lent itself to elaboration and expression on all levels—from philosophy to pseudo-science to magic.

Religions of the West. In Western monotheistic religions it is necessary to distinguish between the role of miracles on the level of popular beliefs and practices and the theory of miracles propounded by the theologians. Belief in a personal, omnipotent Creator who exercises his providence over his creatures implies a concept of miracles as deliberate interventions in the course of events by the same sovereign God who also assures their normal regularity.

The Old Testament and rabbinic Judaism

Judaism. Miracles are taken for granted throughout the Old Testament. God does "wondrous things" according to Psalms, chapter 72, and "great things and unsearchable, marvellous things without number" according to the Book of Job, chapter 5; these things are done in his creation in general and in the history of his people in particular (*e.g.,* the 10 plagues of Egypt and the events of the Exodus). A list of the great wonders done by God is given in Psalms, chapter 136; their purpose is to make his creatures praise him, acknowledge his rule, and "know that I am the Lord." God's wondrous deeds range from the normal regularities of creation to extraordinary interventions that run counter to ordinary experience and thus serve as signs of his greatness and providence in wreaking vengeance on the wicked and giving salvation to his elect.

Later rabbinic Judaism took the occurrence of miracles for granted. It assumed a natural order in which things worked and within which humans were supposed to discharge their duties; thus, to rely on miracles was nothing short of sinful. In special circumstances, however, or in connection with persons of extraordinary saintliness, God would intervene or spectacularly answer their petitionary prayers. It was not so much a matter of suspending as of relativizing nature, the normal course of which was just one possible expression of the divine will. It was only in the Middle Ages and under the influence of Greco-Arabic philosophy that the problem of miracles was systematically discussed on a philosophical and theological level. Normative, rabbinic Judaism, being mainly concerned with doing God's will as revealed in his Law, had little interest in miracles, though it accepted, as a matter of course, the veracity of the miracles recorded in Scripture and in the Talmud (the collection of Jewish lore, legend, and law). On the level of popular piety both magic and the belief in miracles always flourished, especially under the influence of Kabbala, the esoteric, mystical movement within Judaism; the Ḥasidic movement (a pietist movement that arose in eastern Europe in the 18th century) in particular produced a rich crop of beliefs and legends concerning the miraculous virtue—through prayer, intercession, or magical power—of the great Ḥasidic saints and rabbis.

Jesus Christ

Christianity. New Testament accounts of the advent, birth, life, Passion, and Resurrection of Christ include many miracles. Jesus is reported in the Gospels to have performed miracles of diverse kinds: raising the dead, healing the sick, casting out demons, and causing nature miracles, such as the multiplication of loaves and the turning of water into wine at the town of Cana. Unlike the Buddha and Muḥammad, Jesus had an ambiguous attitude toward miracles: on the one hand he performed them as a sign of his mission and of the impending coming of the Kingdom; on the other hand he reproved the desire for wonders and repeatedly forbade the disciples to publicize his miracles, insisting that it was faith alone that worked miracles. In fact, because miracles also could be explained by attributing them to demonic agency, it was ultimately faith that determined the quality and function of the miracle.

Early Christianity developed in the atmosphere of Hellenistic, Greco-Roman culture, which was full of miraculous accounts and legends. These no doubt influenced Christian traditions and forms of devotion, especially as popular religion always hankered after miracles, and—at the conclusion of the Gospel According to Mark—Christ himself had promised the continuance of miracles in his church. In a world in which only a few critical minds doubted the reality of miracles, the similarity of the Christian signs to those reported in pagan legend was attributed to demonic imitation and counterfeit. The problem of distinguishing between the two sources of miracles—because the devil often disguises himself as an angel of light—frequently solicited the attention of theologians and mystics. Whereas for the theologians a miracle was a sign of God's saving presence and design, for the mass of believers it was the manifestation of a sacred power inherent in individual persons, places, and objects.

Medieval theologians—and specifically St. Thomas Aquinas—taught that as all knowledge was derived from sensible facts, so also "a certain degree of supernatural knowledge of the objects of faith" could be brought about "by certain supernatural effects that are called miracles." This doctrine already assumes a system of natural causality that God—though he normally works through the natural law of which he is the author and Creator—can temporarily set aside. It also assumes that—at least in theory, if not always in practice—natural and supernatural effects can be distinguished. Thus, in 1870 the first Vatican Council declared: "If anyone should say that no miracles can be performed, . . . or that they can never be known with certainty, or that by them the divine origin of the Christian religion cannot be rightly proved—let him be anathema." Belief in miracles is thus obligatory in the Roman Catholic Church, although belief in any specific miracle is not necessarily so. Classical Protestantism, however, has confined its belief in miracles to those recorded in Scripture.

The Prophet Muḥammad

Islām. Muslim religion assumes, as a matter of course, that Allāh works miracles and has done so in the past; *e.g.,* through Moses, Solomon, and Jesus but significantly not through the Prophet Muḥammad. According to the Qur'ān, Muḥammad explicitly rejected the idea of proving his vocation by signs and miracles: the Qur'ān itself was the greatest miracle, and he was but a human messenger and preacher of repentance. Nevertheless, subsequent narratives invested his birth and life with superlatively miraculous details.

Muslim popular religion—particularly under Ṣūfī (Islāmic mysticism) influence—abounds in miracles, pilgrimages to the tombs of wonder-working saints, and the like. Dogmatic theology, too, recognizes miracles as facts. The peculiar feature of Muslim theology is that, unlike Christian theology, it did not accept the idea of nature as an entity operating according to fixed laws ordained by the Creator. Because the universe is constantly being recreated by Allāh in successive time atoms, natural regularity is nothing but the regularity of Allāh's habit in recreating the universe. Thus, a miracle is the omnipotent God's departure from his habit but no different, in principle, from the latter. Muslim dogmatics distinguish between miracles (*karāmāt*), with which Allāh surrounds his saints (*awliyāʾ*) as a mark of distinction, and signs (*āyāt,* also *muʿjizāt;* literally, "acts of an overwhelming nature"). The latter are wrought by Allāh to prove the genuineness of his messengers and to overwhelm and reduce to silence their opponents. Such miracles, which deviate from the usual course of things and are of such nature that others cannot produce their like, are Allāh's testimony to the sincerity of his apostles. The problem is nevertheless complicated by the fact that Satan too can perform miracles. Generally speaking, miracles do not play a role in the continued life of orthodox Islām, though they loom large in popular belief and piety.

INTERPRETATION OF MIRACLES

All the more fully developed theologies have formulated a doctrine of miracles in the context of their beliefs regarding

God, the world, the operations of nature, and causality. The emergence of the concept of nature as a closed system functioning in accordance with strict causal laws created problems more than once, but medieval Christian and Jewish thought had no difficulty in maintaining that the order created by God could also be suspended by him.

In classical antiquity. Miracles were denied even in classical antiquity. Thus, Cicero asserted that "nothing happens without a cause, and nothing happens unless it can happen. When that which can happen does in fact happen, it cannot be considered a miracle. Hence, there are no miracles." Cicero qualified this statement, however, by saying that miracle stories may be necessary for the piety of ignorant folk. The 2nd-century pagan philosopher Celsus is less dogmatic in his attacks on Christianity: the Christian miracles are insufficiently attested and most improbable, but, even if they were genuine, they could hardly offset the miracles of the pagan world—*e.g.*, the healings of Asclepius. This was the standard pattern of many religious polemics: miracles as such were not necessarily denied; only those claimed by the adversary were denied. When these could not be denied, they were ascribed to diabolic agency or to the fraudulent practices of priests or occasionally to a misinterpretation of essentially natural phenomena.

Rationalist criticisms

In the 18th and early 19th centuries. Rationalist criticism, although not completely absent in the Middle Ages, became a major factor in the 18th and 19th centuries. David Hume, a British empiricist and a skeptic, in the chapter "On Miracles" in his *Enquiry Concerning Human Understanding* argued that, given the general experience of the uniformity of nature, miracles were highly improbable and that the evidence in their favour was far from convincing. It should be emphasized that Hume, whose criticism led him to a denial of causality, did not dismiss miracles because they were inconsistent with causal law—as many other thinkers did, notably the Deists (those, especially British, who advocated a natural religion). Instead Hume insisted on the probability factor and thus on the importance of assessing historical evidence. Because all Christians agreed that biblical religion and the scheme of salvation set forth in it could be maintained only by stressing prophecy and miracle, there developed a vast body of literature, especially among Protestants, proving the authenticity of the Christian faith on the basis of the miracles recorded in the Bible. For many 18th-century thinkers, however, the vastness and complexity of the order of nature were even more impressive than any alleged exceptions to it. Thus, belief in miracles, although remaining an essential element of faith to good Christians, appeared as sheer superstition to the eyes of the torch-bearers of Rationalist enlightenment.

Criticism of the concept of miracle was articulated in more than one way. There was philosophical and scientific criticism to the effect that miracles were impossible and that even epistemologically (*i.e.*, within the limits of knowledge) the occurrence of a miracle could never be established; at most, these critics maintained, there were merely as yet unexplained natural phenomena. (This view comes close to the religious assertion that faith precedes the experience of a miracle and that the factuality of a miracle can never precede faith.) There was historical and philological criticism, arguing that the actual occurrence of miracles is unsubstantiated and analyzing the growth and evolution of the legends and texts reporting miracles. There was psychological criticism, arguing that some people want to believe in miracles and so produce imaginative creations answering their psychological needs. There was also a type of religious criticism implying that the truly spiritual has no need of miraculous supports. One suggested solution to the problem was the assertion that the term miracle does not describe an objective event but rather a subjective mode of experience. This view of Friedrich Schleiermacher, an early-19th-century Protestant theologian and philosopher, identified miracle with a religious understanding of any aspect of the world.

Liberal and Neo-orthodox Protestantism

In the late 19th and 20th centuries. Later 19th- and 20th-century liberal Protestant thinkers, such as Rudolf Bultmann, a German New Testament scholar, discarded the traditional notion of miracle together with other elements of what they termed the mythological apparatus of the Bible. Many of these liberal theologians sought evidence for Christianity in the moral and religious transformation it brought to people's lives or interpreted the doctrine of salvation in Existential terms. The early decades of the 20th century, however, also witnessed a return to a more orthodox theological climate—as, for example, in the thought of Karl Barth, a Swiss Protestant theologian—and a new readiness to accept miracles as meaningful signs of God's salvific activity. This change of climate coincided with certain developments in science that appeared to question a too rigid and mechanical concept of causal determinism.

Orthodox Jews, Christians, and Muslims still believe in the literal occurrence of the miracles recorded in their scriptures and traditions; Roman Catholics, furthermore, believe in the continued occurrence of miracles, defining them as a direct divine effect upon nature. The liberal attitude—whatever the variations in detail and in sophistication of the explanation—is essentially similar to that propounded by Schleiermacher. (R.J.Z.W./Ed.)

Saint

The phenomenon of saintliness (*i.e.*, the quality of holiness, involving a special relationship to the sacred sphere as well as moral perfection or exceptional teaching abilities) is widespread in the religions of the world, both ancient and contemporary. Various types of religious personages have been recognized as saints, both by popular acclaim and official pronouncement, and their influence on the religious masses (the broad spectrum of those holding various wide-ranging religious beliefs) has been, and is, of considerable significance.

NATURE AND SIGNIFICANCE

Saints are persons believed to be connected in a special manner with what is viewed as sacred reality—gods, spiritual powers, mythical realms, and other aspects of the sacred or holy. The existence of such persons has been a widespread phenomenon throughout the religions of the world. The religious person may have various relationships with the sacred: as seer, prophet, saviour, monk, nun, priest, priestess, or other such personage. In the case of each of these, however, a specific kind of relationship to the holy is involved. Seers, for example, have an inspirational vision of the future; prophets proclaim a revelation; saviours are entrusted with effecting redemption, liberation, or other salvatory conditions; monks and nuns lead religious lives in accordance with ascetic regulations that they generally observe as long as they live. Every one of these religious personages may simultaneously be, or become, a saint, but there is no necessary connection. Sainthood thus implies a special type of relationship to the holy, a relationship that is not automatically obtained by other religious personages through their performance of religious duties or offices.

Relationships with the sacred or holy

The significance of saintly personages is generally based on real or alleged deeds and qualities that became apparent during their lifetimes and continue to exert influence after their deaths. The special character of their feats and qualities of living is believed to arise from an especially close association with a deity or sacred power. In addition to such a relationship, sainthood also requires the existence of a sacral institution that can grant such recognition, or of a popular cult that acknowledges and posits a belief in the saint's special qualities. In institutionalized religions, such as Roman Catholicism, there is a regularized process (called canonization) by which saints are officially recognized. Canonization requires, among other things, proof that the person in question wrought miracles during his or her lifetime. On the other hand, folk belief often recognizes the saintly powers of a living or dead person long before the institutional religion acknowledges him as a saint.

SAINTS IN EASTERN RELIGIONS

Confucianism and Taoism. Confucianism is in the main ethically oriented. Confucius taught that right conduct was

The importance of right conduct

a means of acquiring ideal harmony with the Way (Tao) of Heaven and that the "holy rulers of primal times" were representative examples of such ideal conduct. In the oldest known Chinese historical work, the *Shu Ching* ("Classic of History"), such a ruler, King T'ang (11th century BC), is described as one who "possessed the highest degree of virtue, and so it came to be that he acquired the bright authority of Heaven." Thus, in Confucianism, the saintliness of its holy men lay in ethical perfection, and through the practice of ethical ideals a contact with Heaven (T'ien) was established. Confucius himself serves as an example of a man who was first regarded as a saint because of his deep wisdom and conscientious observance of ethical precepts and was even considered to be "more than human." During the Han dynasty (206 BC–AD 220), Confucius was elevated to a new status: Emperor Kao Tsu offered sacrifice at the Confucian temple, and Emperor Wu proclaimed Confucianism the official ideology of China. The titles duke (AD 1) and king (739) were further tributes to "the perfect sage." During the T'ang dynasty (618–907), sacrifices were regularly offered in Confucian temples, and in 1906 Confucius was declared equal to the Lord of Heaven.

Taoism is oriented toward another kind of sanctity: the attainment of a passionless unity with the Absolute. Chuang-tzu (died *c.* 300 BC), a mystical Taoist sage, speaks of the "pure men of early times" in his work, the *Chuang-tzu,* and characterizes them as such.

Shintō. Shintō, the native Japanese religion, is concerned with the veneration of nature and with ancestor worship; it does not have saints according to the standards of ethical perfection or of exceptionally meritorious performance. According to Shintō belief, every person after his death becomes a *kami,* a supernatural being who continues to have a part in the life of the community, nation, and family. Good men become good and beneficial *kamis,* bad men become pernicious ones. Being elevated to the status of a divine being is not a privilege peculiar to those with saintly qualities, for evil men also become *kamis.* There are in Shintō, however, venerated mythical saints—such as Ōkuni-nushi (Master of the Great Land) and Sukuma-Bikona (a dwarf deity)—who are considered to be the discoverers and patrons of medicine, magic, and the art of brewing rice.

Buddhism. Founded by Siddhārta Gautama, Buddhism developed into three major forms in the course of its more than 2,500-year history: Theravāda ("Way of the Elders"), also called in derogation Hīnayāna ("Lesser Vehicle"); Mahāyāna ("Greater Vehicle"); and, stemming from it, Vajrayāna ("Vehicle of the Thunderbolt"). A belief in saints prevails in all three groups.

Differing concepts of saints in Buddhism

Theravāda Buddhism, claiming strict adherence to the teachings of the Buddha, recognizes as saints (*arhats*) those who have attained Nirvāṇa (the state of bliss) and hence salvation from *saṃsāra* (the compulsory circle of rebirth) by their own efforts. The Buddha himself—having obtained Nirvāṇa ("the destruction of greed, . . . hate, . . . and illusion")—is viewed as the first Buddhist saint. Disciples of the Buddha who reached Nirvāṇa after him also are considered holy men. Furthermore, in early Buddhism, there were also women regarded as holy, including Prajāpatī, the Buddha's aunt and stepmother—whose repeated requests finally caused the Buddha to permit women to enter his order—and his wife Yaśodharā.

Mahāyāna saints

Mahāyāna Buddhism, originating about the beginning of the Christian Era, rejected the Theravāda belief that only monks may attain salvation. In Mahāyāna belief there is a path to redemption for all people, irrespective of their social standing. Salvation and the way to redemption are conceived in terms more liberal than those of Theravāda. Mahāyāna Buddhists believe in an otherworldly paradise that allows for personal existence and in which dwell heavenly Buddhas (those who have attained Nirvāṇa in previous worlds) and *bodhisattvas* ("Buddhas-to-be"). The heavenly Buddhas and *bodhisattvas* are believed to grant grace to sentient beings, so that salvation is no longer acquired by fleeing from the world and giving up worldly professions, but rather by faith (in the sense of trust) in the promise of a saviour deity. Thus, in Mahāyāna Buddhism, the Buddhas and *bodhisattvas* are viewed as the holy ones, the saints, who in compassion, attempt to aid others struggling for salvation. This concept is in striking contrast to the *arhats* of Theravāda Buddhism, who follow the dying Budddha's last words, "Seek *your own* salvation with diligence." The basic altruistic concept of Mahāyāna then is that of the helping *bodhisattva.* Everyone should strive for this ideal in order to save as many fellowmen as possible as a *bodhisattva* and to bring them into the "Greater Vehicle" (Mahāyāna). Hence, the idea of faith in benevolent saints gains prominence in Mahāyāna Buddhism as a theistic religion of salvation. In Japanese Mahāyāna there are patron saints, such as Shōtoku Taishi, the regent who supported the introduction and development of Buddhism in his country in about AD 600, after it had been introduced in AD 552.

Vajrayāna Buddhism, embodying, among other views, Tantrism (a system of magical and esoteric practices), is mainly represented by Tibetan Buddhism. In addition to the innumerable saints of Mahāyāna Buddhism, Tibetan Buddhism also accepts as living saints those who are regarded as incarnations (*tulkus*) of saints, scholars of the past, deities, or demons. The Dalai Lamas, heads of the Tibetan hierarchy, are viewed as reincarnations of Chenre-zi (the *bodhisattva* of mercy, Avalokiteśvara).

Jainism. According to Jain teaching, there were 23 Tīrthaṅkaras (saintly prophets or proclaimers of salvation) before Mahāvīra Vardhamāna, the 6th-century-BC Indian religious leader after whom Jainism was named. Today they are venerated as saints in temples containing their images. Veneration of the Holy Tīrthaṅkaras is viewed in terms of purifying the devotee morally, as these saints are but examples for the Jainas and not actually objects of a cult.

Hinduism. Hinduism in a wider sense encompasses Brahmanism, a belief in the Universal Soul, Brahman; in a narrower sense it comprises the post-Buddhist, caste-ordered religious and cultural world of India. The Indian religions are by and large mystical in character; hence, even in early Hinduism ascetics were highly honoured. Mysticism generally starts with ascetic practices as a means of eliminating a desire for worldly existence.

Significance of asceticism in the Hindu concept of saint

In later Hinduism, when the ascetics continued to be revered by the masses as *sādhus* (saints, or "good ones") and *yogis* (ascetic practitioners), the concept of the *avatāra* (the idea of the incarnation of a divine being in human form) served to interpret the existence of holy men. By means of this concept it was, and still is, possible to consider living and dead saints as incarnations of a deity and also to incorporate saints of other religions into the Hindu world of belief. Thus Jesus Christ, for instance, is regarded as an *avatāra* of the god Vishnu (Viṣṇu), and the Hindu saint Rāmakrishna is considered to be an *avatāra* of the god Śiva.

SAINTS IN WESTERN RELIGIONS

Ancient Greek religion. The ancient heroes of Greek religion may be regarded as saints. One basis for belief in heroes and the hero cult was the idea that the mighty dead continued to live and to be active as spiritual powers from the sites of their graves. Another source of the cult of heroes was the conception that gods were often lowered to the status of heroes. One of the best known heroes is Heracles, who became famous through his mighty deeds. In Greek religion the numinous (spiritual) qualities of a person lay in such heroic deeds.

Zoroastrianism and Parsiism. Zoroastrianism includes the veneration of Fravashis—*i.e.,* preexistent souls that are good by nature, gods and goddesses of individual families and clans, and physical elements. According to Zoroastrian belief, humans are caught up in a great cosmic struggle between the forces of good, led by Ahura Mazdā (Wise Lord), and the forces of evil, led by Angra Mainyu, or Ahriman, the Evil Spirit. In the battle between Asha (Truth) and Druj (Lie) the Fravashis may correspond to the saints of Roman Catholicism, who can be called upon for aid in times of trouble.

Judaism. The cult of saints in terms of veneration was not a part of the monotheistic religion of Israel. Saintliness,

however, was an ideal that many hoped to exhibit. The model of a pious person is depicted in the righteous one of Psalm 5, "his delight is in the law of the Lord, and on his law he meditates day and night." In the Hellenistic period (*c.* 300 BC—*c.* AD 300), when many Jews were susceptible to foreign religious influences, the Ḥasidim (the "pious" ones) segregated themselves from the others, holding fast to the faith of their fathers.

The Ḥasidim

The concept of the Ḥasidim gained new significance in the 18th century when Israel ben Eliezer, called Ba'al Shem Ṭov, or "Master of the Good Name," started the modern movement called Ḥasidism. As opposed to the Orthodox Israelite religion with its emphasis on rationalism, cultic piety, and legalism, Ba'al Shem Ṭov stood for a more mystically oriented form of Judaism.

Christianity. Jesus and his disciples did not speak of saints; but during the period (1st to early 4th century) in which they were persecuted, Christians began to venerate the martyrs as saints. They believed that the martyrs, being sufferers "unto death" for Christ, were received directly into heaven and could therefore be effective as intercessors for the living. By the 3rd century the veneration of martyr saints was already common.

In the Nicene Creed (AD 325) the early church called itself the "communion of saints." Here, however, the word "saint" has the broader meaning of "believer" rather than being applied strictly to a holy person or numinous personality worthy of veneration. In the 10th century a procedure of canonization (official recognition of a public cult of a saint) was initiated by Pope John XV. Gradually, a fixed process was developed for canonization by the pope, requiring that the person must have led a life of heroic sanctity and performed at least two miracles.

Saints in the Roman Catholic Church are venerated—but not worshipped—because of their spiritual and religious significance and are believed to be the bearers of special powers. Because of a belief in the powers of the saints, their relics are regarded as efficacious. In the Eastern Orthodox Church saints also are venerated, but the process of canonization is less juridical and not always ecumenical. In some Protestant churches (Lutheran and Anglican) saints are recognized, but not venerated as in the Roman Catholic and Orthodox.

Islām. Islām is a rigorously monotheistic religion, strictly prohibiting any kind of "conjunction" (*i.e.*, affiliation, or consortship) to Allāh. Thus the concept of sainthood was rejected. Yet even here a variegated belief in holy men arose because of the demands of popular religion. Over against the one distant God, whose almighty power and whose role as a strict judge was emphasized repeatedly, there emerged a desire for intercessors. These were found in saintly men who were believed to be endowed with charismatic powers (*karāmāt*), allowing them to go miraculously from one place to another far away; to wield authority over animals, plants, and clouds; and to bridge the gap between life and death. The Prophet Muḥammad (died AD 632) had negated the existence of saints, but the piety of the masses "canonized" holy men while they were still living. After they died, cults of devotion arose at the sites of their graves, and pilgrimages to such sites were believed to aid the believer in acquiring help and blessing.

Saintly powers of Muslim holy men

MODES OF RECOGNITION

The bases of recognition. The basic motive for the belief in and veneration of saints is, primarily, the recognition by people of religious persons whom they view as holy. In order for a religious personage (*e.g.*, prophet) to be recognized as a saint, it is necessary that other people see in him the aura of holiness. The holiness recognized in him may be an impersonal sacred or spiritual power—which is often perceived in quite insignificant persons—and is believed to be present even in the bones and other material relics of a recognized holy person after his death. Religious personalities also are believed to possess a personal holiness, either bestowed upon them by divine grace or acquired through ascetism and moral discipline. Such sanctity reveals itself in the power to perform miracles.

Recognition of forms of holiness

The highest form of holiness in a holy person is reflected in the interpretation of that person as an incarnation of divine reality or as the possessor of godly nature. Divine qualities are perceived in such a person, and through him, such as the Logos (divine Word, or Reason) in Jesus.

Popular recognition. Popular recognition of saints arises out of a predilection of the religious masses (those who maintain popular belief, or folk belief, along with beliefs officially promulgated) to grasp the supernatural in that which is believed to be unusual and uncommon—*i.e.*, in the miraculous event. Thus, the religious masses long for those who can perform wonders that are awe awakening and satisfy their desire for the miraculous and mysterious.

Besides the desire for miracles, there is another basic requirement of the masses, especially within monotheistic religions: the yearning for a superhuman being in human form. The one abstract God who is believed to be present everywhere and capable of helping everybody and everything is too unperceptual and remote for the average religious person. There is a tendency among the religious masses to split up the deity into many numinous beings that fulfill the desires of the people. The religious masses often have polytheistic tendencies. The term "dear saints," as the holy ones are called in Roman Catholicism, expresses an emotional relationship to those near, benevolent, heavenly, or spiritual powers that are the heirs to the ancient ethnic and patron deities of pre-Christian times.

In the course of their histories, and as they expand, the great universal religions (*e.g.*, Christianity, Buddhism, and others) incorporate ever more people with their particular folk beliefs. As their numbers grow and their influence increases in the religious communities, the indigenous peoples retaining many earlier folk beliefs form the majority and their inclinations prevail. Because their behaviour patterns generally remain constant, their religious forms are preserved. Occasionally, religious reform movements arise within the organized mass religions. Such movements attempt to restore what is believed to be the original form of the respective religions and often turn against a belief in and veneration of saints, regarding such forms of religiosity as degenerate. This was the case in the 16th-century Protestant Reformation and also in the Wahhābiyah movement, an 18th-century reform movement in Islām.

Theological interpretations of popular recognition. In monotheistic religions the belief in saints in its popular form generally contradicts orthodox teaching. Such religiosity is usually opposed and rejected or else reinterpreted in view of its ineradicability. If the latter is the case, the orthodox interpretation given the cult of saints in order to justify it is a theological construction. In Roman Catholicism, for instance, church doctrine makes a distinction between veneration (*veneratio, douleia*) and adoration (*adoratio, latreia*). Veneration is defined as a proper attitude toward saints, whereas adoration is applicable only in connection with God. The veneration of images as practiced especially in the Eastern Orthodox Church is explained similarly. The Roman Catholic Church also teaches that the saints are representatives of God's grace on earth and that they are completely subject to his will. The vestigial remains of polytheistic beliefs and practices connected with the veneration of saints are thus theologically, though not popularly, eliminated.

Theological distinctions of forms of veneration

Similar interpretations of the belief in saints in a monotheistic religion serve to justify an existing cult. The people themselves are hardly influenced by such interpretations, however. According to many scholars, the differentiation between *douleia* (veneration) and *latreia* (worship), or between *veneratio* (veneration) and *adoratio* (adoration), has little meaning for the masses. In practice, they observe their cult of saints quite in accordance with polytheistic devotion toward gods. The supplications actually directed to the saints in the various religions can hardly be distinguished from prayers to deities, even though the saints are theologically regarded as mere intercessors having special access to God, and the answer to prayer is considered as coming from God alone. From the perspective of scholars of comparative religion, however, beings to whom prayers are dedicated are gods.

Forms of cults. The form of a cult of saints can be categorized as either *indirect* or *direct*. An indirect cult form

Indirect and direct cultic forms

involves the veneration of objects that stand in a magical relationship with the respective saint. In this connection there can be a veneration of the saint's relics. Such religious practices are to be understood in terms of spiritual power. Numinous power is viewed as issuing from the saint; and it is believed to be acquired by veneration or, in practice, mainly by touching (or kissing) the object itself. Another indirect cult form is the veneration of the image of the saint. According to primitive belief, there is a magical connection between the image and the original, which is itself holy. A common and widespread custom is the depositing of votive offerings, dedicated to certain saints, at holy places—temples, churches, shrines, or chapels where the supplicant can be certain of their direct presence and aid. This custom is of ancient origin—*e.g.*, the votive offerings dedicated to the healing god Asclepius in the museum of Epidaurus (Greece). This practice is still to be found in present-day popular belief in Greece or at Roman Catholic places of pilgrimage.

In these forms of indirect cult, then, saints are venerated through the medium of concrete objects. In direct veneration, on the other hand, the saint himself is addressed in invocation and praise. According to popular belief, such direct worship is most effective at the place of the predominant presence of the respective saint. The idea of pilgrimage is always based upon such a belief in the localized presence of numinous power.

TYPES AND FUNCTIONS OF SAINTS

Saints as moral examples. A classical illustration of the saint who is distinguished by his virtue is St. Francis of Assisi. Giving up a life of extravagance, he began in 1209, together with several friends, to actualize his ideal of the imitation of Christ by leading a life of poverty. For St. Francis, three virtues constituted the preconditions of true divine vision: poverty, ascetic chastity, and humility.

An example of a similar kind of saintliness is reflected in the person of the Indian leader and reformer Mahatma Gandhi (1869–1948). In his life, devoted to the acquiring of freedom for India, he also lived according to three ideals. The first was *satyāgraha,* holding fast to the truth with all the powers of the spirit. Gandhi's second basic principle was *ahiṃsā,* which is to be understood not only in the negative sense of "not killing" but also positively as a renunciation of the self and an indulgence in "kind actions" toward all beings. His third ideal was *brahmacarya,* which often is rendered too narrowly as chastity; it is the ascetic way of life that Gandhi followed as a saint and as a statesman, hence receiving boundless veneration by the masses.

Saints as prophets and reformers. Many prophets and prophetic reformers form a second group of saints. One prophet in early Christianity was Paul, who is honoured as a saint by Roman Catholics, Eastern Orthodox, and Protestants. He was a most powerful spiritual personality, decisively and significantly involved in the development of Christianity from a Jewish sect to a world religion.

The Tibetan reformer Tsong-kha-pa belonged to a completely different world from that of St. Paul. Originally, he did not want to be an innovator but only a renewer of old religious patterns. He was mainly concerned with the restoration of the discipline and the development of the Lamaistic cult. His fame grew, and owing to his activity many monasteries were founded. The "Yellow Hat" sect was established by him. According to legend, Tsong-kha-pa was taken up to heaven before the eyes of the people. This accounts for the veneration he received, and still receives, by the Tibetan people.

Theological teachers as saints. Often numbered among the saints are certain religious personalities whose significance lies in their work as illuminating interpreters of religious tradition or as proponents of a new view of the divine or the eternal. An example from Indian religions is the great teacher (*ācārya*) Śaṅkara, the representative of Advaita (the teaching of the nonduality of divine reality). When he died at the age of 32, a short and outwardly uneventful life lay behind him. Yet even today the personality and work of Śaṅkara continue to determine the intellectual and religious life of India. Equally significant in the Christian West, and specifically in the Roman Catholic Church, is Thomas Aquinas, a Dominican scholar. Although first disputed, his work finally received general recognition, and he became recognized as the *doctor communis* ("general teacher") of the Roman Catholic Church. His significance lies in his encompassing and methodically clear theological and philosophical system, in which he reconciled the views of the ancient Greek philosopher Plato with those of his student Aristotle, antiquity with Christianity, knowledge with faith, and nature with grace. He was proclaimed a saint in 1323. (G.Me./Ed.)

BIBLIOGRAPHY

General. HEINRICH EMIL BRUNNER, *Dogmatik,* vol. 1 (1946; Eng. trans., *Dogmatics,* vol. 1, *The Christian Doctrine of God,* 1949), advocates the primacy of Scripture over tradition; OWEN CHADWICK, *From Bossuet to Newman: The Idea of Doctrinal Development* (1957), an excellent survey of the gradual shift from "the classical consciousness" of identity in doctrine (Bossuet) to a "historical consciousness" of growth and continuity (Newman); ADOLF VON HARNACK, *Lehrbuch der Dogmengeschichte,* 3rd ed., 3 vol. (1893; Eng. trans., *History of Dogma,* 7 vol., 1900, reprinted 1961), a massive exposition of the thesis that Christian dogma represents the process of Hellenization of the original Gospel, hence a deviation; JOHN HENRY NEWMAN, *An Essay on the Development of Christian Doctrine,* new ed. (1878), a classical statement of the emergence of the historical consciousness within the Catholic tradition; J. PELIKAN, *Development of Christian Doctrine: Some Historical Prolegomena* (1969), an important statement of the interaction of Scripture and tradition in the formation of Christian doctrines and dogmas; FREDERICK J. STRENG, *Understanding Religious Man* (1969), an excellent summary of the common elements in religious experience, including those relating to doctrine and dogma; R.C. ZAEHNER, *Concordant Discord: The Interdependence of Faiths* (1970), helpful insights as to the various ideas of authority in the major religions of the world.

Creation. *Cosmogonic myths:* CHARLES H. LONG, *Alpha: The Myths of Creation* (1963), gives examples of various types of cosmogonic myths from different cultures. For ancient Near Eastern myths, see *Ancient Near Eastern Texts Relating to the Old Testament,* ed. by JAMES B. PRITCHARD, 3rd ed. with suppl. (1969). JOHANNES PEDERSEN, *Israel,* 4 vol. (Eng. trans. 1926–40), is a cultural-religious study that shows the relationship between creation myth, land, and kinship system. For the nature and structure of myths and symbols, see ERNST CASSIRER, *Philosophie der symbolischen Formen,* 4 vol. (1953–56; Eng. trans., *The Philosophy of Symbolic Forms,* 3 vol., 1953–55); and JOAN O'BRIEN and WILFRED MAJOR, *In the Beginning: Creation Myths from Ancient Mesopotamia, Israel, and Greece* (1982).

The development and structure of Greek myths: JOHN BURNET, *Early Greek Philosophy,* 4th ed. (1930, reprinted 1963), is a well-written interpretation of the pre-Socratic myths of creation. ARNOLD EHRHARDT, *The Beginning* (1968), shows the common structure of the cosmologies of the Gospel According to John and pre-Socratic thinkers.

Christian doctrine: For a theological history of the Christian doctrine of creation in its variety and continuity, see JAROSLAV PELIKAN, *Development of Christian Doctrine* (1969), *The Christian Tradition* (1971), and *Historical Theology: Continuity and Change in Christian Doctrine* (1971). JOHN MACQUARRIE, *Principles of Christian Theology* (1966), presents a structural and systematic analysis of the elements of Christian theology, showing how the doctrine of creation fits into theological systems.

Islām: DE LACY O'LEARY, *Arabic Thought and Its Place in History,* rev. ed. (1939, reprinted 1963), deals with the internal and external sources of Arabic philosophy and cosmology. SEYYED HOSSEIN NASR, *An Introduction to Islamic Cosmological Doctrines* (1964), explicates a tradition in Arabic thought that expresses creation in symbolic and cosmological images.

Zoroastrianism: Several Zoroastrian myths and doctrines of creation are found in R.C. ZAEHNER, *The Dawn and Twilight of Zoroastrianism* (1961).

Chinese philosophy: ARTHUR F. WRIGHT (ed.), *Studies in Chinese Thought* (1953), brings together 10 essays on various aspects of Chinese thought; most valuable is DERK BODDE, "Harmony and Conflict in Chinese Philosophy," pp. 19–80. For a history of Chinese philosophical speculation as it relates to cosmogony and cosmology, see FUNG YU-LAN, *A History of Chinese Philosophy,* 2nd ed., 2 vol. (1952–53).

Indian philosophy: Speculations about creation in the various schools of Indian philosophy can be found in SURENDRANATH DAS GUPTA, *A History of Indian Philosophy,* 5 vol. (1922–55). ALAIN DANIELOU, *Le Polythéisme hindou* (1960; Eng. trans., *Hindu Polytheism,* 1964), is a description and interpretation of the gods of Hinduism in relationship to their philosophical

meaning. T.R.V. MURTI, *The Central Philosophy of Buddhism* (1955), is an explication of the Mādhyamika system of Buddhist philosophy that denies creation.

Comparative works: HAJIME NAKAMURA, *Ways of Thinking of Eastern Peoples* (1964), is a comparative work showing the similarities and contrasts between Indian, Chinese, Tibetan, and Japanese modes of thought especially as they concern creation. C.F. VON WEIZSACKER, *The Relevance of Science: Creation and Cosmogony* (1964) deals with the evolution of thought about creation from myth to scientific theory. CHARLES HARTSHORNE and WILLIAM REESE (eds.), *Philosophers Speak of God* (1953), explores the rational bases for several conceptions of God and creation in Eastern and Western thought.

Eschatology. *General:* M. ELIADE, *Le Mythe de l'Éternel retour* (1949; Eng. trans., *The Myth of the Eternal Return*, 1954), an internationally recognized standard work; T. ROSZAK, *The Making of a Counter Culture* (1969), a discussion of the formation of secular messianic movements in modern society; W.D. WALLIS, *Messiahs: Their Role in Civilization* (1943); GAYRAUD S. WILMORE, *Last Things First* (1982), a scholarly discussion of eschatology as it has existed in various cultures and religions.

Biblical: R.H. CHARLES, *Eschatology: The Doctrine of a Future Life in Israel, Judaism and Christianity* (1899, reprinted 1963), a comparison of ideas; O. CULLMANN, *Christus und die Zeit*, 3rd ed. (1962; Eng. trans., *Christ and Time*, 1962); R.K. BULTMANN, *History and Eschatology* (1957), an existentialist interpretation of eschatology; P.S. MINEAR, *The Christian Hope and the Second Coming* (1954), an integral biblical view; R.J. ZWI WERBLOWSKY, "Messianism in Jewish History," *Journal of World History*, 11:30–45 (1968), with a bibliography; and N.R.C. COHN, *The Pursuit of the Millennium*, 2nd ed. (1961; rev. paperback ed., 1970).

Theological and philosophical: M. SCHMAUS, *Von den letzten Dingen* (1948), the best Roman Catholic treatise on eschatology; N. BERDYAEV, *The Beginning and the End* (1957, orig. pub. in Russian, 1947), Russian Orthodox philosophy of religion; J.A.T. ROBINSON, *In the End, God* (1968), an introduction to this subject; P. TEILHARD DE CHARDIN, *Le Phénomene humain* (1956; Eng. trans., *The Phenomenon of Man*, 1959), a discussion combining eschatology with the theory of evolution; J. MOLTMANN, *Theologie der Hoffnung*, 8th ed. (1964; Eng. trans., *Theology of Hope*, 1967), on the beginning of the ecumenical theology-of-hope movement.

Angels and demons. C. JOUCO BLEEKER and GEO WIDENGREN (eds.), *Historia Religionum: Handbook for the History of Religions*, vol. 1, *Religions of the Past* (1969), and vol. 2, *Religions of the Present* (1971), contains helpful sections on the role of angels and demons in chapters on the various religions, as well as a very usable bibliography. J.B. NOSS, *Man's Religions*, 4th ed. (1969), contains useful sections on angels and demons. GUSTAV DAVIDSON, *A Dictionary of Angels, Including the Fallen Angels* (1967); and ROSSELL H. ROBBINS, *The Encyclopedia of Witchcraft and Demonology* (1959), are Western-oriented. R.C. ZAEHNER, *The Dawn and Twilight of Zoroastrianism* (1961), has excellent sections on the role of angels and demons in Zoroastrianism and their relationship to Hindu spiritual beings. ROBERT M. GRANT, *Gnosticism and Early Christianity*, 2nd ed. (1966), contains useful sections relating angelic and demonic figures of Judaism, Christianity, and Zoroastrianism to Gnostic speculation. See also JEFFREY B. RUSSELL, *Satan: the Early Christian Tradition* (1981).

Salvation. For recent and comprehensive studies of the subject, S.G.F. BRANDON, *Man and His Destiny in the Great Religions* (1962), provides extensive documentation and bibliographies; important aspects of salvation are specially dealt with in his *History, Time and Deity* (1965) and *The Judgment of the Dead* (1967). S.G.F. BRANDON (ed.), *The Saviour God* (1963), comprises 15 essays by specialists in the major religions. A valuable and well-documented study of the subject in Hebrew, Greco-Roman religions, and early Christianity may be found in T. KLAUSER (ed.), *Reallexikon für Antike und Christentum*, vol. 5, col. 54–219 (1964). ADOLF VON HARNACK'S monumental *Lehrbuch der Dogmengeschichte*, 3rd ed., 3 vol. (1893; Eng. trans., *History of Dogma*, 7 vol., 1900, reprinted 1961), traces the development of Christian soteriology. L.W. GRENSTED, *A Short History of the Doctrine of the Atonement* (1920, reprinted 1962), is a reliable concise guide. Aspects of salvation in the religions concerned are treated in the following books: R.C. ZAEHNER, *Hinduism*, 2nd ed. (1966) and *The Dawn and Twilight of Zoroastrianism* (1961); E.J. THOMAS, *The History of Buddhist Thought*, 2nd ed. (1951); and A.J. WENSINCK, *The Muslim Creed* (1932).

Providence. No general introduction to the subject written from the point of view of the science of religion exists, but useful articles are found in some specialized encyclopaedias, such as *Hastings' Encyclopaedia of Religion and Ethics* (1919). Further information has to be gathered from monographs about specific problems related to Providence, *e.g.*, JOHN BOWKER, *Problems of Suffering in the Religions of the World* (1970).

Revelation. R.C. ZAEHNER, *At Sundry Times* (1958), a sympathetic approach by an accomplished scholar who finds anticipations of Christian revelation not only in Judaism but also in Hinduism, Buddhism, and Zoroastrianism; J.H. WALGRAVE, *Un salut aux dimensions du monde*, trans. from the Dutch (1970), an apologetically oriented work that attempts to bring out the distinctive qualities of the Christian view of revelation in comparison with Buddhism, Hinduism, and Islām.

Primitive religion: MIRCEA ELIADE, *Traité d'histoire des religions* (1948; Eng. trans., *Patterns in Comparative Religion*, 1958), a discussion of hierophanies, myths, and symbols as pertinent to the theme of revelation; G. VAN DER LEEUW, *Phänomenologie der Religion* (1933; Eng. trans., *Religion in Essence and Manifestation*, 2 vol., 1963), a phenomenological approach influenced by Rudolf Otto and others.

Christianity: A.R. DULLES, *Revelation Theology: A History* (1969), a survey of Catholic and Protestant views; J. BAILLIE, *The Idea of Revelation in Recent Thought* (1956), a sketch of trends in 20th-century Protestant theology; CARL F. HENRY, *God, Revelation and Authority*, 6 vol. (1976–83), an evangelical's argument for the infallibility of biblical revelation.

Islām: A.J. ARBERRY, *Revelation and Reason in Islam* (1957), a concise and learned treatment of the medieval controversies; K. CRAGG, *The Call of the Minaret*, pt. 2, pp. 33–171 (1956), a very objective presentation of Muslim faith and piety, including some discussion of the doctrine of revelation.

Hinduism: K.S. MURTY, *Revelation and Reason in Advaita Vedānta* (1959), an exposition and evaluation of Śaṅkara's position in the light of modern Western philosophy; R.C. ZAEHNER, *Hindu and Muslim Mysticism* (1960), on the love relationship to God in Bhakti and Ṣūfism.

Buddhism: W.L. KING, *Buddhism and Christianity: Some Bridges of Understanding* (1962), an objective comparison between Christianity and Theravāda Buddhism, with a good discussion of the revelatory role of the Buddha.

Judaism: A.J. HESCHEL, *God in Search of Man*, pt. 2, pp. 167–278 (1956), a presentation of modern Judaism by a prominent rabbinic scholar.

Covenant. W. BEYERLIN, *Herkunft und Geschichte der ältesten Sinai traditionen* (1961; Eng. trans., *Origins and History of the Oldest Sinaitic Traditions*, 1966); G.E. MENDENHALL, "Covenant," *Interpreter's Dictionary of the Bible*, vol. 1, pp. 714–723 (1962); R.C. DARNELL, *Idea of Divine Covenant in the Qur'an* (1970); D.J. MCCARTHY, *Treaty and Covenant* (1963); "Egyptian and Hittite Treaties" in J.B. PRITCHARD (ed.), *Ancient Near Eastern Texts Relating to the Old Testament*, 2nd ed., pp. 199–206 (1955); D.R. HILLERS, *Covenant: The History of a Biblical Idea* (1969); JONATHAN BISHOP, *The Covenant* (1982), a look at covenants in the Old and New Testaments and an attempt to apply the concept to today's world.

Prophecy. *General:* G. HOLSCHER, *Die Prophteten* (1914), a classic; A.J. HESCHEL, *The Prophets* (1962), a theological comparison between Israelite and non-Israelite prophets; J. LINDBLOM, *Prophecy in Ancient Israel* (1962), a good introduction to the phenomenological, psychological, and theological problems of prophecy; R.B.Y. SCOTT, *The Relevance of the Prophets*, 2nd ed. (1968).

Prophecy in the ancient Middle East and Israel: A. GUILLAUME, *Prophecy and Divination Among the Hebrews and Other Semites* (1938), a standard work; D.R. HILLERS, *Treaty-Curses and Old Testament Prophets* (1964); A.L. OPPENHEIM, *Ancient Mesopotamia: Portrait of a Dead Civilization* (1964); R.E. CLEMENTS, *Prophecy and Covenant* (1965), a valuable study of the available prophetical texts from Mari thus far; N.K. GOTTWALD, *All the Kingdoms of the Earth* (1964), on prophets and politics; E. HAMMERSHAIMB, *Some Aspects of Old Testament Prophecy from Isaiah to Malachi* (1966), dealing with the Canaanite, cultic, and historical background; A.R. JOHNSON, *The Cultic Prophet in Ancient Israel*, 2nd ed. (1962); J. PEDERSEN, *Israel*, 4 vol. (1926–40), a classic on religious life and institutions; H. RINGGREN, *Israelite Religion* (1966).

Prophecy in Christianity: L. HARTMAN, *Prophecy Interpreted* (1966); H.A. GUY, *New Testament Prophecy* (1947); G. FRIEDRICH, "Prophets and Prophecies in the New Testament," *Theological Dictionary of the New Testament*, vol. 6, pp. 828–861 (1968); S. UMEN, *Pharisaism and Jesus* (1963).

Prophecy in Islām: T. ANDRAE, *Mohammed: The Man and His Faith* (Eng. trans. 1956); S. FUCHS, *Rebellious Prophets* (1965); A. GUILLAUME, *Islam*, new ed. (1963); P.K. HITTI, *Islam: A Way of Life* (1970); W. MONTGOMERY WATT, "Muḥammad," *The Cambridge History of Islam*, vol. 1, pp. 30–56 (1970).

Prophetic movements and figures in Eastern and primitive religions: I. HORI, *Folk Religion in Japan: Continuity and Change*, ed. by J.M. KITAGAWA and A.L. MILLER (1968); E.R.

and K. HUGHES, *Religion in China* (1950); B.G.M. SUNDKLER, *Bantu Prophets in South Africa*, 2nd ed. (1961); M. WEBER, *The Religion of China*, trans. by H.H. GERTH (1968).

Miracle. GUSTAV MENSCHING, *Das Wunder im Glauben und Aberglauben der Völker* (1957), the best and most complete treatment of the subject; ROBERT M. GRANT, *Miracle and Natural Law in Graeco-Roman and Early Christian Thought* (1952), a description of Hellenistic attitudes and beliefs. BENEDICTA WARD, *Miracles and the Medieval Mind: Theory, Record, and Events, 1000–1215* (1982), a study of the attitudes toward reports of miracles by the church and by other institutions.

Saint. H. RINGGREN and A.V. STROM, *Die Religionen der Völker* (1959; Eng. trans. from the 3rd Swedish ed., *Religions of Mankind Today and Yesterday*, 1967), gives information on the significance of saints. The cult of saints is dealt with in connection with the phenomenology of religion by G. MENSCHING, *Die Religion* (1959). See also W.J. BURGHARDT, *Saints and Sanctity* (1965). The characteristics and the actions of holy men in non-Christian religions are treated in R.A. NICHOLSON, *The Mystics of Islam* (1914, reprinted 1963); W.T. DE BARY *et al.* (comps.), *Sources of Chinese Tradition* (1960); and G. VON GRUNEBAUM, *Medieval Islam*, 2nd ed. (1953). In the realm of Christianity, P. MOLINARI, *I Santi e il loro culto* (1962; Eng. trans., *Saints: Their Place in the Church*, 1965), gives information concerning the veneration of saints in folk piety. The veneration of saints in the Eastern Church is canvassed by D. ATTWATER, *Saints of the East* (1963).

Dogs

The dog is one of the two most ubiquitous and popular domestic animals in the world (the cat is the other). For more than 12,000 years the dog has lived with humans as a hunting companion, protector, object of scorn or adoration, and friend. The dog has evolved from similar (that is, undifferentiated) fur-bearing animals into more than 400 distinct breeds. Human beings have played a major role in creating dogs that fulfill distinct societal needs. Through the most rudimentary form of genetic engineering, dogs were bred to accentuate instincts that were evident from their earliest encounters with humans. Although details about the evolution of dogs are uncertain, the first dogs were hunters with keen senses of sight and smell. Humans developed these instincts and created new breeds as need or desire arose.

Dogs are regarded differently in different parts of the world. Western civilization has given the relationship between human and dog great importance, but, in some of the developing nations and in many areas of Asia, dogs are not held in the same esteem. In some areas of the world, dogs are used as guards or beasts of burden or even for food, whereas, in the United States and Europe, dogs are protected and admired. In ancient Egypt during the days of the pharoahs, dogs were considered to be sacred.

Characteristics of loyalty, friendship, protectiveness, and affection have earned dogs an important position in Western society, and in the United States and Europe the care and feeding of dogs has become a multibillion-dollar business.

All dogs belong to the family Canidae, along with their relatives—wolves, jackals, and foxes. They are members of the mammalian order Carnivora, or "Flesh Eaters." Although there are more than 400 different dog breeds, all dogs belong to a single species, *Canis familiaris*.

Dogs have played an important role in the history of human civilization and were among the first domesticated animals. They were important in hunter-gatherer societies as hunting allies and bodyguards against predators. When livestock were domesticated about 7,000 to 9,000 years ago, dogs served as herders and guardians of sheep, goats, and cattle.

Although many still serve in these capacities, dogs are increasingly used for social purposes and companionship. Today, dogs are employed as guides for the blind and disabled or for police work. Dogs are even used in therapy in nursing homes and hospitals to encourage patients toward recovery. Humans have bred a wide range of different dogs adapted to serve a variety of functions. This has been enhanced by improvements in veterinary care and animal husbandry.

For coverage of related topics in the *Macropædia* and *Micropædia*, see the *Propædia*, section 313, and the *Index*. This article is divided into the following sections:

Origin and history of dogs 440
- Ancestry
- Domestication
- Origin of breeds

Physical traits and functions 441
- General characteristics
- Reproduction

Behaviour 443
- Territory and range
- Barking
- Behavioral development
- Breed-specific behaviour

Dogs as pets 444
- Selection
- Nutrition and growth
- Training
- Other maintenance concerns
- Ailments

The breeds 445
- Sporting dogs
- Hounds
- Terriers
- Working dogs
- Herding dogs
- Toys
- Non-Sporting dogs
- Breed standards

Related canids 449
- Wolves
- Coyotes
- Foxes
- Jackals
- Other wild canids

Bibliography 450

ORIGIN AND HISTORY OF DOGS

Ancestry. Paleontologists and archaeologists have determined that about 60 million years ago a small mammal, rather like a weasel, lived in the environs of what are now parts of Asia. It is called *Miacis*, the genus that became the ancestor of the animals known today as canids: dogs, jackals, wolves, and foxes. *Miacis* did not leave direct descendants, but doglike canids evolved from it. By about 30 to 40 million years ago *Miacis* had evolved into the first true dog—namely, *Cynodictis*. This was a medium-size animal, longer than it was tall, with a long tail and a fairly brushy coat. Over the millennia *Cynodictis* gave rise to two branches, one in Africa and the other in Eurasia. The Eurasian branch was called *Tomarctus* and is the progenitor of wolves, dogs, and foxes.

It is believed that the early dogs dating from about 12,000 to 14,000 years ago came from a small strain of gray wolf that inhabited what is now India. Thereafter, this wolf—known as *Canis lupus pallipes*—was widely distributed throughout Europe, Asia, and North America. It is also

possible that some of the dogs of today descended not from the wolf but rather from the jackal. These dogs, found in Africa, might have given rise to some of the present native African breeds.

No matter what their origins, all canids have certain common characteristics. They are mammals that bear live young. The females have mammary glands, and they suckle their offspring. The early breeds had erect ears and pointed or wedge-shaped muzzles, similar to the northern breeds common today. Most of the carnivores have similar dental structures, which is one way paleontologists have been able to identify them. They develop two sets of teeth, deciduous ("baby") teeth and permanent teeth.

Canids walk on their toes, in contrast to an animal like the bear, which is flat-footed and walks on its heels. Dogs, like most mammals, have body hair and are homeothermic—that is to say, they have an internal thermostat that permits them to maintain their body temperature at a constant level despite the outside temperature.

Fossil remains suggest that five distinct types of dogs existed by the beginning of the Bronze Age (about 4500 BC). They were the mastiffs, wolf-type dogs, sight hounds (such as the Saluki or greyhound), pointing dogs, and herding dogs.

First encounter with humans

Domestication. It is uncertain when the first dog became a companion of humans, but it is likely that wild canids were scavengers near tribal campsites at the same time that ancient humans discovered a hunting partner in the animals that ventured close by. In ancient Egypt, dogs were thought to possess godlike characteristics. They were pampered by their own servants, outfitted with jeweled collars, and fed the choicest diet. Only royalty was permitted to own purebred dogs, and upon the death of a ruler his favourite dog was often interred with him to protect him from harm in the afterlife.

Illustrations of dogs dating from the Bronze Age have been found on walls, tombs, and scrolls throughout Europe, the Middle East, and North America. Often the dogs are depicted hunting game with their human counterparts. Statues of dogs guard the entrances to burial crypts. In many cases these dogs clearly resemble modern canines. Such relics are indelible testimony to the importance that humans have given to the dog throughout the ages.

Origin of breeds. Once it became evident that dogs were faster and stronger and could see and hear better than humans, those specimens exhibiting these qualities were interbred to enhance such attributes. Fleet-footed sight hounds were revered by noblemen in the Middle East, while in Europe powerful dogs such as the mastiff were developed to protect home and traveler from harm.

As society changed and agriculture—in addition to hunting—became a means of sustaining life, other breeds of dogs were developed. Herding and guarding dogs were important to farmers for protecting their flocks. At the same time, small breeds became desirable as playthings and companions for noble families. The Pekingese in China and fragile breeds such as the Chihuahua were bred to be lapdogs. The terrier breeds were developed, mainly in England, to rid granaries and barns of rodents. Pointing and retrieving breeds were selected for special tasks related to aiding the hunter to find and capture game. Many breeds are extremely ancient, while others have been developed as recently as the 1800s.

PHYSICAL TRAITS AND FUNCTIONS

General characteristics. Dogs come in a wide range of shapes and sizes. It is difficult to imagine that a large Great Dane and a tiny poodle are of the same species, but they are genetically identical with the same anatomic features. All dogs have 78 chromosomes, or 39 pairs of chromosomes (humans have 23 pairs). One member of each pair comes from each parent.

Teeth. Dogs have two sets of teeth. Twenty-eight deciduous teeth erupt by six to eight weeks of age, and by the time puppies are six to seven months old these deciduous teeth are all replaced by 42 adult teeth. The permanent teeth include incisors, which are used to nip and bite; canines, which tear and shred flesh; and premolars and molars, which shear and crush. In short, a dog's teeth serve as weapons and as tools for cutting or tearing food. The canines are the upper and lower fangs for which the dog family was named. As in most carnivores, the teeth are high-crowned and pointed, unlike the broad, grinding teeth of many herbivorous animals.

The teething process can be difficult for puppies. Their gums hurt and become swollen, they may lose their appetites, and they may have mild intermittent diarrhea.

Digestive system. Dogs rarely chew their food. Once the food is taken into the mouth, it is gulped or swallowed and passed through the esophagus into the stomach, where digestive enzymes begin to break it down. Most of the digestion and absorption of food takes place in the small intestines with the aid of the pancreas and the liver. The pancreas secretes enzymes needed for regulating the digestive process. As in humans, the pancreas produces insulin and glucagon, both of which are necessary for the regulation of glucose. The liver is the largest internal organ in the body. It has six lobes (whereas the human liver has only two). The liver is responsible for many essential life-preserving functions. It helps digestion by producing bile, which aids in the absorption of fat. The liver also metabolizes protein and carbohydrates, and it excretes toxins from the bloodstream. In addition, it manufactures major blood-clotting agents. Because the liver performs all these vital functions, liver disease can be a major problem in dogs.

Skeletal structure. The skeletal frame of the dog consists of 319 bones. If a dog's tail is docked or absent at birth, there obviously are fewer bones in the skeleton. The muscles and tendons of a dog are similar to those of a human; however, a dog's upper body muscles bear half the weight of the entire body and are better developed than a human's. The weight distribution between the front and the rear of the dog are relatively equal.

Dogs are running animals, with the exception of those bred specifically for different purposes. For instance, the bulldog, with its large head and short, "bowed" legs, cannot be called a creature born to chase game. Most dogs, however, are well equipped to run or lope over long distances, provided that they are physically conditioned for such activities. The construction of the shoulder and pelvic bones and the way they articulate with the leg bones and the spine allow most breeds to trot, run, or gallop with ease. Certain breeds have distinct gaits that have been genetically selected by humans. The German shepherd dog is known for its "flying trot." The extreme extension of the front and rear legs causes the dog to appear as if it were soaring, although one foot always remains on the ground. Another unique gait is that of the greyhound. This dog was bred for great bursts of speed, and its most comfortable gait is the gallop. The spine is unusually flexible, allowing the dog to contract and extend its four legs in unison, whereby all four feet are off the ground at the same time.

Bone construction

Other breeds also have unique features. The Afghan hound was bred to chase game over long distances in rocky terrain. Its structure permits great flexibility through the hip joints and lower back, enabling the dog to turn quickly in a small area. The dachshund, by contrast, is long and low with short legs. This dog was bred to hunt badgers underground, and its shape allows it to enter subterranean tunnels in search of its prey.

Although most breeds no longer follow the pursuits for which they were originally bred, their instincts remain strong, and their structure still allows them to perform their specific tasks.

Senses. Dogs have the same five senses as humans. However, some are more highly developed, and others are deficient compared with those of humans. Dogs' sense of smell is by far the most acute and is immeasurably better than that of humans. Dogs are used for such tasks as tracking missing persons, digging underground, and tracing toxic substances, such as gases, that are undetectable by humans. Dogs can detect drugs, explosives, and the scents of their masters. Not all canine noses are the same, however. Some breeds, such as the German shepherd and the bloodhound, have much more keenly developed olfactory senses than others. One would not choose a short-nosed breed, such as the pug, to engage in tracking.



A dog's keen sense of smell helps a state trooper search for explosives at an airport.
Michael Grecco/Stock, Boston

Even in short-nosed breeds, however, the olfactory centre is relatively highly developed. It is arranged in folds in order to filter smells from the incoming air. Some rescue dogs are trained to follow a scent on the ground, and others are trained to scent the air. Both are able to distinguish one person from another even after a considerable passage of time. Hunting dogs—such as pointers, retrievers, and spaniels—are trained to scent birds and can distinguish one variety of bird from another.

The dog's sense of taste is poorly developed compared with that of humans. If forced to live on their own, dogs will eat almost anything without much discrimination.

Dogs possess an acute sense of hearing. Aboriginal breeds had large, erect and very mobile ears that enabled them to hear sounds from a great distance in any direction. Some modern breeds have better hearing than others, but they all can detect noises well beyond the range of the human ear. Dogs are able to register sounds of 35,000 vibrations per second (compared with 20,000 per second in humans), and they also can shut off their inner ear in order to filter out distracting sounds.

The eyesight of a dog is not as keen as its sense of smell, and it is generally thought that dogs have poor colour perception. Some breeds, such as the Saluki and the Afghan hound, were developed to chase game by sight over long distances, and these dogs can see well enough to detect any movement far on the horizon. Dogs can generally see better in poor light than humans but not as well in bright light. They have a wider field of vision than humans because their eyes are set further toward the sides of their heads, but they are not as adept at focusing on objects at close range or at judging distances. Dogs have a third eyelid, a membrane that protects the eyeball from irritants and is sometimes visible in front of the eye.

Dogs are sensitive to touch, the fifth sense, and use this sense to communicate with one another and with their human counterparts. Learning where to touch a dog is an important part in either stimulating or relaxing it and is useful in training a puppy or bonding with an adult dog.

Coats. There are three basic types of hair: short (as on a pointer or Doberman pinscher), medium (as on an Irish setter or Siberian husky), and long (as on a chow chow or Maltese). Within these categories there are also coarse and fine types of hair. Dogs come in a wide variety of colours, but in many breeds colour selection is an important consideration, as is the colour distribution on the dog.

Most dogs shed their coats seasonally. This is a natural occurrence that depends in large measure on the amount of available daylight. In the fall as days become shorter, a dog's coat will grow thicker and longer. In the spring the dog will begin to shed its coat, and it will take longer for the coat to grow in over the summer. Temperature influences the amount of body coat a dog grows. Dogs living in warm climates all year long rarely grow hair coats as thick as those living in colder areas, although this will affect the body coat and the amount of protective undercoat more than the topcoat or the length of furnishings on the belly, ears, and tail.

Grooming is an important part of touch to a dog and can be a pleasurable and relaxing means of relating to it. The dog's coat forms a barrier between the environment and the skin. Grooming the coat enhances the dog's beauty and well-being and gives the owner the chance to evaluate the general health of the dog.

Reproduction. *Sexual maturity.* There is some variation in the age at which dogs reach sexual maturity. Small breeds appear to mature faster than large ones, which usually cycle later. It is not uncommon for large-breed females to come into heat for the first time at more than 1 year of age, although 8 to 9 months is the norm. Dogs are sexually mature between 6 months and 1 year but are not socially mature until they are about 2 years of age. Females first cycle anywhere from 6 to 18 months of age and approximately twice a year thereafter. The only exception is the African basenji, which cycles annually, bearing one litter a year.

Reproductive cycle. The heat cycle of the female lasts from 18 to 21 days. The first stage is called proestrus. It begins with mild swelling of the vulva and a bloody discharge. This lasts for about 9 days, although it may vary by 2 or 3 days. During this phase the bitch may attract males, but she is not ready to be bred and will reject all advances. The next phase is the estrus. Usually the discharge decreases and becomes lighter, almost pink, in colour. The vulva becomes very enlarged and soft, and the bitch will be receptive to the male. This stage may last 3 or 4 days or as long as 7 to 11 days. The female may be receptive a day or two past the time when she would still be fertile. In order to be sure that the breeding is taking place at the optimum time, vaginal smears and blood tests can be done by a veterinarian beginning before estrus and through the estral phase.

Female reproductive cycle

At about the 14th day, or whenever estrus ends, the final, or luteal, stage of the cycle begins; this stage is called diestrus. The discharge becomes redder, the vulva returns to its normal size, and the bitch will no longer accept the male for mating. When all signs of discharge and swelling are absent, the heat is complete. The diestrus stage lasts 60 to 90 days (if no pregnancy has occurred) or until the bitch gives birth. She then enters anestrus, which is the time frame between the end of the last cycle and the beginning of the next proestrus.

Canine males are always fertile from the onset of their sexual adolescence, usually after six months of age. Larger-breed males may take a few months longer to become sexually mature. Males are usually promiscuous and are willing to mate with any available female.

Males produce far more sperm than is needed to impregnate the ova that are released during estrus. Small-breed bitches usually produce small litters. Two or 3 puppies in a breed such as a Yorkshire terrier is considered the norm. Large-breed litters can have as many as 10 or 12 puppies, although the normal bitch can suckle up to 8 at a time.

Gestation and whelping. The normal gestation period is 63 days from the time of conception. This may vary if the bitch has been bred two or three times or if the eggs are fertilized a day or two after the mating has taken place. Eggs remain fertile for about 48 hours. Sperm can live in the vaginal tract for several days. In order to determine if a bitch is pregnant, a veterinarian can manually palpate her abdomen at about 25 days after breeding. Ultrasound also can be done at that time. At about 40 days X rays will confirm pregnancy.

Most bitches whelp normally. However, the large-headed, short-bodied breeds and the toy breeds often must undergo cesarean sections in order to deliver live puppies.

Reproductive capacity. Both males and females are fertile well into their advanced age. It is generally considered best for the bitch to be bred for the first time upon maturity but not before her second or third heat cycle, depending on her age at the first. Because small breeds mature more quickly, they can be bred at an earlier age

than large breeds. A bitch will have less difficulty in conceiving and carrying a litter if she is bred before the age of five. As she becomes older, litter size generally decreases. After the age of seven, bitches are likely to have small litters and experience problems in delivering the puppies. Veterinarians feel that bitches generally should not be bred after that age.

Males can be bred as long as they are fertile, although with age the motility and quantity of sperm decrease.

BEHAVIOUR

The dog is a social creature. It prefers the company of people and of other dogs to living alone. It is, therefore, considered by animal behaviourists to be a pack animal. In this respect it is similar to its distant relative the wolf. As a result of millennia of selective breeding, the dog has been adapted to live with people. Seminal studies of dog behaviour conducted in the 1950s and '60s showed, however, that dogs raised without human contact at an early age retain their inherent instincts and prefer relationships with other dogs over associations with people.

Territory and range. Both dogs and wolves are territorial animals. Wolf packs, because of their need to hunt game, claim large territories as their own, whereas dogs claim their territories based on the limitations of their owners. Male wolves and dogs mark their territorial boundaries by urinating and rubbing their scent on the ground or on trees to warn other animals of their presence.

When on neutral ground, that which is not considered by either dogs or wolves to be their home territory, strangers greeting each other will go through formal rituals of sniffing, marking, tail wagging, and posturing. Unless they are claiming the same prey or are engaged in courting the same female, such interactions are usually terminated by each going its own way. Females will attack strangers in neutral territory to protect their young, however.

Types of barks

Barking. Both dogs and wolves have a repertoire of barks, growls, and howls that are identifiable among themselves and to humans who have studied their vocabulary. Dog owners can determine by certain sounds whether their pet is playful, warning of a stranger nearby, fearful, or hurt. One of the earliest signs that puppies are becoming social and independent creatures within the litter are the yips and barks that they make while playing with one another. Dogs, unlike wolves, will growl if cornered or fearful. Certain breeds of dogs, notably hounds, have been bred to enhance the howling instinct when they are on the trail of game. Some of the northern breeds, such as the Siberian husky, howl rather than bark. At the other end of the spectrum, the basenji does not bark but rather emits a yodeling sound when it is happy.

Behavioral development. Canine behaviour is a combination of instinct and environment. Dogs are born with certain innate characteristics that are evident from birth.

Puppy development

Puppies are born blind and deaf, totally dependent on the dam for warmth and nourishment. The dam will instinctively suckle and protect her young, often keeping other dogs and all but the most trusted people away from the whelping box. Between 10 and 14 days after birth, the eyes and ear canals open, and the puppies begin to move actively around their nest. As they grow, they become more curious and start to investigate their surroundings independently. The dam will begin to leave them alone briefly. During this phase they relate most intensely to their littermates and dam and may become unhappy at being removed from their familiar surroundings. This stage of development lasts about 20 days and is the first of four critical periods.

Beginning at three weeks of age, the most adventurous puppies will seek ways to get out of the whelping box and will start to investigate the larger world. At this age puppies are receptive to human contact, which is essential if they are to bond with people when they become adults. Dogs left alone from four weeks on will never reach their full potential as pets and will often become independent and more difficult to train than those accustomed to close human contact from an early age. At the same time, during the period between three and seven weeks, it is important that puppies socialize with their littermates and dam. This is when the dam weans her puppies, first by regurgitating some of her own food and then by not allowing her puppies to nurse as often as they would like. At about four weeks of age, puppies can be offered solid food in the form of a soft gruel.

Individual socialization of each puppy in a litter can begin at six weeks of age. This is when puppies begin to be more receptive to handling and attention.

The third critical period in a puppy's development is from 7 to 12 weeks. It has been shown in studies undertaken at various breeding kennels that this is the best age to form human-dog relationships. Attachments formed during this period will affect the attitude of the dog toward humans and toward its acceptance of direction and learning. During this period the pack instinct, which has played such an important role in the puppy's early development, can be transferred to humans. At this time environment becomes a vital part of the dog's education and training. This is when a human can most easily establish dominance over the dog, becoming the "leader of the pack." At this age a dog will accept a submissive role more readily than at any other time in its life. Learning comes most readily at this age. Puppies taught basic commands, even if they are not reinforced for several months, will remember them and respond if they are taught during this critical age.

The fourth critical stage in a puppy's development is between 12 and 16 weeks. At this age the puppy will declare its independence from its mother and will become increasingly daring in its forays from the familiar. Puppy training can begin during this period, and it is a time of rapid physical and mental growth. The permanent teeth begin to emerge at this time, which is often a painful and distractive process. Puppies need to chew during this period, and, if they are not provided with appropriate teething toys, they will use any available hard object, such as furniture. Puppies at this age may be less willing to cooperate or respond to new commands.

A dog's personality continues to develop during its entire maturing process and will undergo radical changes while the dog matures sexually and physically. Dogs mature sexually earlier than they do emotionally. Their personalities develop more slowly than their bodies, much like humans but unlike wolves, whose personalities and sexuality develop more harmoniously.

At about seven or eight months many puppies tend to go through a period of anxiety. They are insecure, frightened of strangers, and will appear timid. If this is not an inherited trait, it will disappear within a few months. If it is inherited, that condition will remain and may become accentuated with time.

Breed-specific behaviour. There are distinctive breed-typical personalities that have been developed through generations of selection for certain traits. By roughly grouping dogs according to the work they were bred to do, it is possible to determine the type of temperament a dog might have at maturity. Differences in breed personalities can be seen at an early age. Sporting dogs will generally be adventurous, following their noses wherever scents lead them, but will respond enthusiastically to calls from familiar humans. Hounds generally tend to be more aloof and independent, inclined to scout the territory on their own and follow a scent or a movement; they are not as interested in human interaction as the bird dogs are.

Working and herding dogs have more business-like dispositions. They tend to evaluate situations and set about their tasks. Collie puppies have been known to herd children, ducklings, or each other in an instinctive manifestation of their birthright. Guarding dogs tend to be protective of their territories, even at an early age. Such dogs as the Maremma or the kuvasz, which are bred to guard flocks, are placed with the sheep from the time they are puppies in order to reinforce their basic protective instincts. Collies and Akitas are known for their strong sense of loyalty. Terriers, bred to chase and catch rodents, have a tendency to be extremely active, lively, and feisty as puppies, traits that continue into adulthood. Newfoundlands are renowned for lifesaving instincts.

Breed specificity also affects how well dogs adapt to new surroundings or to new owners. Such things cannot be

taught to dogs. They are innate—part of a dog's instinctive behaviour—and are often breed-specific, although mixed breeds have been known for unique instincts as well.

DOGS AS PETS

The companionship between humans and dogs is not a new phenomenon. However, in modern society most dogs are owned as pets, not because of the work they were bred to do. Many breeds, such as the toy dogs, were developed precisely to be pets. All of the diverse breeds and mixed breeds have unique traits and appeal to different kinds of people.

Acquiring a dog is a major decision, because the dog becomes totally dependent on its owner for its care and welfare. This responsibility continues throughout the life of the dog. Thus, the initial decision should be based on a serious consideration of whether one's lifestyle truly lends itself to owning a dog—that is, whether a dog would be an asset rather than a liability.

Selecting a breed

Selection. The next consideration is the selection of a particular type of dog. Many people want a purebred dog because they like the appearance or the personality, and they are assured that the puppy they buy will grow up to look like the breed it represents. Others find that a mixed breed will do just as well, and there are many shelters, humane societies, and rescue groups that harbour dogs in need of homes.

No matter what kind of dog a person chooses, it is essential that it be a healthy animal. When evaluating a puppy or an adult dog, several features will help determine the physical condition of the animal. The dog should appear friendly and outgoing. Puppies in particular should exhibit curiosity and a tail-wagging enthusiasm. They should not hang back or appear timid or frightened. Eyes should be bright and shiny with no discharge, and the inner eyelids ought to be smooth and pink. Ears should be clean-smelling and free of debris. Gums must be pink and firm, except in the case of chow chows and shar-peis, whose gums and tongue are black. The skin should feel warm and dry to the touch. Clammy skin or the presence of reddened patches, crusts, scales, or parasites are indicative of problems that could be both external and internal. The hair coat ought to be clean and sweet-smelling. The dog should be in good form and build, but not obese or so thin that the ribs and hipbones show.

People buying purebred dogs should know the distinctive characteristics of the breed they have chosen, so that they can ask the breeder proper questions and have some means of evaluating the quality of the dog they are purchasing. Many purebred dogs have hidden genetic problems of which good breeders are aware. Many of these problems can be controlled by careful breeding, but the purchaser must know—through reading about the breed and talking to fanciers—what questions to ask. Mixed-breed dogs also can have hidden genetic problems, but there is no way to determine what they might be or whether they will eventually affect the dog in an adverse manner.

Great strides are being made in veterinary research to identify genetic defects and thereby assist breeders to select the best breeding stock. By eliminating from their gene pool those dogs with genetic abnormalities, breeders can help ensure that the breed remains healthy and viable.

Nutritional requirements

Nutrition and growth. Puppies need three basic things in order to thrive: good nutrition, warmth, and companionship. Puppies need to eat three or four times a day from the time they are weaned until they are about six months old. Thereafter they can be fed twice a day until maturity and once daily after that. However, many dog owners, especially those with large breeds, feed twice a day throughout the dog's life (this does not mean feeding more than the required daily amount, but it is a more balanced method of feeding).

Puppies need twice an adult dog's maintenance requirements of energy and nutrients for proper growth from the time they are weaned until they reach about half of their expected mature weight. There should be steady growth on a weekly basis, but there should be no excess fat around the abdomen. Puppies grow best if they remain at a suitable weight without becoming obese. Overweight puppies are candidates for crippling bone diseases if they are too heavy during the critical growing months. On the other hand, feeding too little will result in poor growth and lack of energy.

Adult dogs burn fewer calories than do puppies or young and active adults. Therefore, they need to eat less in order to maintain optimum weight and activity.

Dogs that work require extra nutrients. For instance, sled dogs need to be fed a diet that is much higher in calories, one with a ratio of fat, protein, and carbohydrates very different from the diet of more sedentary dogs. Owners may have to experiment with different types of food to determine which are best suited to their dogs.

There are three basic types of commercially produced dog foods: canned, dry, and semimoist. Predominant ingredients of most of these include corn, wheat, barley, rice, or soy meal, in combination or alone. Commercial dog foods also include a meat such as beef, lamb, chicken, or liver, or meat by-products. It is important to read the labels to determine the proportions of each and the amounts of proteins, carbohydrates, fats, and vitamins and minerals contained.

Sleep is almost as important as nutrition for puppies. A warm, quiet place for them to rest is essential for normal growth. Puppies will usually play vigorously and then suddenly fall asleep. Their need for sleep decreases as they grow into adulthood, but dogs spend a great deal of their time sleeping when they are not stimulated to activity.

Exercise requirements

All dogs need exercise, some more than others. Achieving good health and sound temperament demands that dogs be given the opportunity for regular stimulating exercise. Puppies should be allowed to run at will without restraint and without being pushed beyond their limits. As dogs mature, jogging or walking on a lead can be introduced, but any forced exercise should be withheld until the dog is fully grown. The most common cause of a dog's destructive behaviour in the house is lack of exercise. Behavioral problems such as tail chasing, chewing, and excessive barking and whining can in most cases be traced to confinement for long periods of time without respite. The ability to provide adequate exercise is one of the most important considerations that prospective dog owners must face before acquiring a puppy. Exercise, however, does not mean allowing the dog to run at large. Dogs ought to be supervised at all times when outside: they either should be accompanied by owners using a lead or have a securely fenced area in which to play.

The term companion animal means that dogs need company. They are happiest when allowed to be an integral part of the household. Puppies thrive and learn when they are included in the household routine at an early age. Training becomes easier when the unique bond between human and dog is strengthened from the beginning.

Training. Puppies learn by watching, but their instincts guide how readily they will learn certain basic requirements. A dog bred to guard the home will be less likely to run off following a scent than a bird dog bred to hunt game. On the other hand, a guarding breed will need direction concerning who is "acceptable" and who is not, whereas a retriever will befriend everyone. Knowledge of what a dog was bred to do is useful when trying to train it to be an acceptable companion.

There are many theories about how to train a dog to be a happy and willing companion, but certain principles apply to all methods. The dog must understand what is expected. It has to be praised for doing well. Punishment for an infraction should be immediate and appropriate to the act. The dog must be able to associate the punishment with the crime. Consistency and kindness bring the best results in training. Most dogs will accept domination readily, but there are some, usually males, who will challenge that authority. This is dangerous behaviour and must be stopped at an early age. Good training must be sensible, and commands should be enforceable.

Other maintenance concerns. Dogs need regular care from the time they are born. In addition to a balanced diet, grooming is an important part of maintaining good health. Care of the ears, coat, and nails on a weekly basis gives owners an opportunity to examine their pets and to



Clumber spaniel.



Labrador retriever.



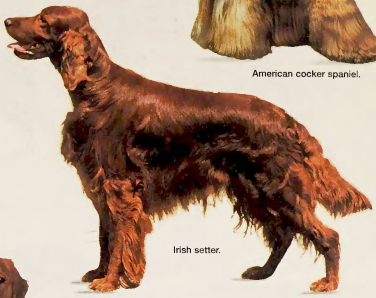
Brittany spaniel.



Chesapeake Bay retriever.



American cocker spaniel.



Irish setter.



English springer spaniel.



Golden retriever.

German shorthaired pointer.



Pointer on point.





English cocker spaniel.

Plate 1: (Top left) © Paddy Cutts/Animals Unlimited, (upper middle left, centre left) © Kent & Donna Dannen, (bottom left, centre right) © R.T. Wilhite/Animal Photography, (lower middle left) © Ron Kimball, all other photographs by © Sally Anne Thompson/Animal Photography

Sporting Breeds



Weimaraner.



Vizsla.

Hounds



Borzoi.



Beagle.



Basenji.

Plate 2: (Centre, lower middle right, bottom right) © Kent & Donna Dannen, (upper middle right) © R.T. Wilbie/Animal Photography, all other photographs by © Sally Anne Thompson/Animal Photography



Whippet.



Saluki.



Norwegian elkhound.



Basset hound.



Bloodhound.



Afghan hound.

Hounds



Irish wolfhound.



Black and tan coonhound.



Greyhound.



Dachshund (standard).

Terriers



Scottish terrier.



Cairn terrier.



Airedale.



Skye terrier.

Plate 3: (Lower middle left, top right, upper middle right centre) © Kent & Donna Dannen. (bottom centre) © R.T. Willis/Animal Photography; all other photographs by © Sally Anne Thompson/Animal Photography



Parson Jack Russell terrier.

Fox terrier (smooth).

Soft-coated wheaten terrier.



West Highland white terrier.

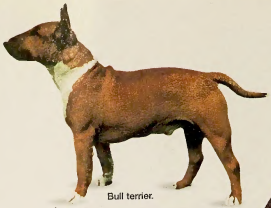




Terriers



Kerry blue terrier.



Bull terrier.



Border terrier.



Bedlington terrier.



Miniature schnauzer.



American Staffordshire terrier.

Working Dogs



Rottweiler.



Akita.

Doberman pinscher.



Siberian husky.

Boxer.

Alaskan Malamute.



Plate 4: (Top left, centre right, lower middle right) © Sally Anne Thompson/Animal Photography, (top centre, centre middle left) © Paddy Cutts/Animals Unlimited, (bottom left) © R.T. Willbie/Animal Photography; all other photographs by Stuart & Donna Dannen

Working Dogs



Great Dane, ears natural (left) and cropped (right).



Bullmastiff.



St. Bernard.



Newfoundland.



Samoyed.



Bernese mountain dog.

Herding Dogs

Plate 5: (Top left, upper middle left) © Ron Kimball, (lower middle left) © R.T. Wittie/Animal Photography, (top right, upper middle right, bottom centre) © Kent & Donna Dannen, (bottom left, lower middle right) © Sally-Anne Thompson/Animal Photography, (bottom right) © Paddy Cutts/Animals Unlimited



Cardigan Welsh corgi.



Pembroke Welsh corgi.



Border collie.

Herding Dogs



Australian shepherd.



Collie.



German shepherd (Alsatian).



Bearded collie.



Bouvier des Flandres.



Old English sheepdog.



Puli.



Shetland sheepdog.



Australian cattle dog.



Belgian sheepdog.



Chihuahua, long-coat (left) and smooth-coat (right).



Shih tzu.



Maltese.



Pekingese.



Chinese crested (hairless).



Cavalier King Charles spaniel.



Papillon.



Pug.



Pomeranian.



Yorkshire terrier.

Plate 7: (Top left, bottom center) © Paddy Cutts/Animals Unlimited, (bottom left, upper middle center) © Kent & Donna Dannen; all other photographs by © Sally Anne Thompson/Animal Photography.

Non-sporting Breeds



Boston terrier.



Chow chow.

Plate 8: (Top left) © Sally Anne Thompson/Animal Photography, (upper middle left, top right, upper middle right, lower middle centre) © Kent & Donna Dannen, (lower middle left, bottom left, upper middle centre) © Paddy Cutts/Animals Unlimited, (lower middle right, bottom right) © Ron Kimball



Poodle (standard).



Keeshond.



Chinese shar-pei.



Lhasa apso.



Bulldog.

Bichon frise.





Dalmatian.

Schipperke.



spot any potential illness. Ears should be cleaned regularly and nails kept trimmed. Brushing should be part of a dog's weekly or even daily routine. Dogs with long or thick coats will need more frequent brushing than shorthaired varieties in order to loosen dead hair and prevent skin irritations or infection.

Regular veterinary care is important to a dog's health. Puppies usually are vaccinated against the most virulent diseases, starting at six weeks of age. A series of three or four vaccinations against distemper, hepatitis, parainfluenza, leptospirosis, and parvovirus are given three weeks apart. At three months of age puppies can be inoculated against rabies. Booster vaccinations are given annually thereafter, except for rabies shots, which may be administered every two or three years, depending on the region. Routine vaccination procedures have succeeded in reducing, and in some areas eliminating, diseases that formerly killed half of all puppies born.

In many areas veterinarians recommend that dogs be tested annually for heartworm disease and be given a preventative. This should be administered throughout the dog's life as long as it resides in a region where and when this parasite is prevalent.

Ailments. Fleas and ticks are sources of irritation and disease in every climate of the world (with the possible exception of the Arctic). Regular bathing and grooming helps to keep these and other external parasites under control. Treatment of the animal and its environment are essential to eliminate these pests. In some areas this is a yearlong process, whereas in other climates it is a seasonal problem.

Common ailments

Internal parasites are a common cause of sickness, especially in puppies. There are many kinds of worms that invade the intestinal tract, resulting in listlessness, loss of blood and subsequent anemia, poor hair coat, and occasionally death. Many of these parasites are found in dirt and are ingested or get into the bloodstream through the skin of the dog. Effective veterinary remedies are available for the animal, but it is important to determine through fecal examination or blood tests exactly what type of parasite is present. Puppies should be examined about every three months, and adults need to be examined annually.

Dogs are susceptible to many of the same illnesses that afflict humans. Cancer, respiratory ailments, allergies, arthritis, and certain forms of heart disease are all found in dogs. Some illnesses have a breed predilection, whereas others occur in all pure and mixed breeds. Large- and giant-breed dogs, such as Irish setters, St. Bernards, bloodhounds, and Great Danes, are prone to a condition known as gastric dilatation volvulus (GDV). This disease causes the stomach to twist in the abdominal cavity, cutting off the blood supply and filling the stomach with gas. GDV is always a medical emergency and must be treated as soon as the first symptoms appear. Early warnings may be restlessness, unsuccessful attempts to vomit or defecate, swelling of the abdomen, or distention of the rib cage.

Large breeds also are at risk for an orthopedic problem in which the hip joint does not develop properly. This is called hip dysplasia and is considered to be a polygenetic condition. It is a progressive disease in which the malformation of the hipbones causes arthritic changes, lameness, and pain. Some breeds are also at risk of developing elbow dysplasia and other problems of the bones and joints. Dogs built with long, low bodies, such as dachshunds, often develop spinal injuries or malformations of the spinal column.

Heart problems

Dogs do not suffer from high cholesterol or from the life-threatening circulatory illnesses that afflict humans, but certain breeds are predisposed to malformations of the heart muscle and valves. Some of these are surgically correctable, while others are not. In addition, heartworm and other parasites may affect the heart and circulatory system.

Dogs are as much at risk of contracting cancers as people are. The treatment is often the same. Cancers most often seen in dogs involve osteosarcomas, mammary tumours, and lymphomas. Veterinary research is at the forefront of the development of new treatments for cancers in the hope that new methods for combating them in humans will be found in the process.

Eye diseases, many of which are hereditary, also are found in dogs. Dogs are subject to cataracts, glaucoma, and retinal diseases, all of which can cause blindness. Treatments in dogs are not as successful as in humans, but dogs appear to adjust to vision loss very well as long as they are kept in familiar surroundings. Their keen sense of smell helps them to get around, although they must be protected from sudden falls and unforeseen dangers. Many canine ocular problems of a hereditary origin are difficult to eradicate because they do not appear in some breeds until the dogs are five or six years old. Nonetheless, genetic research to identify dogs that are carriers or that will develop eye problems has made significant strides since it began in the 1970s.

Breeds with large, protruding eyes, such as the Pekingese or the pug, are susceptible to eye irritations and corneal lacerations. These must be attended to promptly to avoid serious damage to the eye.

Dogs with dropped ears—the basset hound is an extreme example—are prone to diseases of the ear canal. Moisture becomes trapped in the ear, producing yeast infections. Such parasites as ear mites thrive in the ear canal, causing a dark, malodorous exudate. Frequently, the dog is uncomfortable and scratches the ears or rubs the ears along the ground or on the furniture. Most ear problems can be cured with proper medication. If problems are left unattended, the ear canal will develop ulcerations that are painful and difficult to treat.

THE BREEDS

There are approximately 400 separate breeds of purebred dogs worldwide. A purebred dog is considered to be one whose genealogy is traceable for three generations within the same breed. National registries, such as the American Kennel Club (AKC) in the United States, the Canadian Kennel Club, the Kennel Club of England, and the Australian National Kennel Council, maintain pedigrees and stud books on every dog in every breed registered in their respective countries. The Foxhound Kennel Stud Book, published in England in 1844, was one of the earliest registries. Other countries also have systems for registering purebred dogs. The AKC represents an enrollment of more than 36 million since its inception in 1884, and it registers approximately 1.25 million new dogs each year.

In the 1800s those interested in the sport of dogs developed a system for classifying breeds according to their functions. The British classification, established in 1873 and revised periodically by the Kennel Club of England, set the standard that other countries have followed, with some modifications. British, Canadian, and American classifications are basically the same, although some of the terminology is different. For example, Sporting dogs in the United States are Gundogs in England. Utility dogs in England are Non-Sporting dogs in the United States and Canada. Not all countries recognize every breed.

Early classification systems

The United States recognizes seven classifications, called groups (encompassing more than 130 breeds), whereas the English and Canadians have six groups (the American system divides the Working group into two groups: Working dogs and Herding dogs). The groups recognized by the AKC are:

Sporting dogs. These are dogs that scent and either point, flush, or retrieve birds on land and in water. They are the pointers, retrievers, setters, spaniels, and others, such as the vizsla and the Weimaraner.

Hounds. These also are hunting dogs but much more various than the Sporting dogs. There are scent hounds and sight hounds. They are a diverse group, ranging from the low-slung dachshund to the fleet-footed greyhound. However, they all are dedicated to the tasks for which they were bred, whether coursing over rough terrain in search of gazelles, such as the Afghan hound or the Saluki, or going to ground after badgers, like the dachshund. Hounds such as beagles, basset hounds, harriers, foxhounds, and coonhounds run in packs, while others, such as Afghan hounds, borzois, pharaoh hounds, and Salukis, course alone. The Hound group also includes the Petit Basset Griffon Vendéen, the otterhound, the Rhodesian ridgeback, which was bred to hunt lions in Africa, and

Dog Breeds and Their Places of Origin

breed	height (inches)		weight (pounds)	
	dogs	bitches	dogs	bitches
SPORTING				
American water spaniel (United States)	15–18	same	30–45	25–40
Brittany (France)	17½–20½	same	30–40	same
Chesapeake Bay retriever (United States)	23–26	21–24	65–80	55–70
Clumber spaniel (France)	19–20	17–19	70–85	55–70
Cocker spaniel (England)	15	14	24–28	same
Curly-coated retriever (England)	25–27	same	70–80	same
English cocker spaniel (England)	16–17	15–16	28–34	26–32
English setter (England)	25	24	40–70	same
English springer spaniel (England)	20	19	49–55	same
Field spaniel (England)	18	17	35–50	same
Flat-coated retriever (England)	23–24½	22–23½	60–70	same
German shorthaired pointer (Germany)	23–25	21–23	55–70	45–60
German wirehaired pointer (Germany)	24–26	not under 22	60–70	same
Golden retriever (Scotland)	23–24	21½–22½	65–75	55–65
Gordon setter (Scotland)	24–27	23–26	55–80	45–70
Irish setter (Ireland)	27	25	70	60
Irish water spaniel (Ireland)	22–24	21–23	55–65	45–58
Labrador retriever (Canada)	22½–24½	21½–23½	60–75	55–70
Nova Scotia duck tolling retriever (Canada)†	17–21	same	37–51	same
Pointer (England)	25–28	23–26	55–75	45–65
Sussex spaniel (England)	13–15	same	35–45	same
Tahl-Tan bear dog (Canada)†	maximum 15	same	maximum 15	same
Vizsla (Hungary)	22–24	21–23	49–62	same
Weimaraner (Germany)	25–27	23–25	70–85	same
Welsh springer spaniel (Wales)	18–19	17–18	35–45	same
Wirehaired pointing Griffon (Holland)	22–24	20–22	50–60	same
HOUNDS				
Afghan hound (Afghanistan)	approximately 27	approximately 25	60	50
American foxhound (United States)	22–25	21–24	60–70	same
Basenji (Egypt)	17	16	24	22
Basset hound (France)	14	same	40–60	same
Beagle (England)	2 varieties, 13 and 15	same	18–30	same
Black and tan coonhound (United States)	25–27	23–25	NA	NA
Bloodhound (Italy)	25–27	23–25	90–110	80–100
Borzoi (Russia)	at least 28	at least 26	75–105	55–85
Dachshund (miniature; Germany)	smaller than standard	same	11 or under	same
Dachshund (standard; Germany)	7–10	same	16–32	same
English foxhound (England)	23	same	55–75	same
Greyhound (Egypt)	27–30	same	65–70	60–65
Harrier (England)	19–21	same	48–60	same
Ibizan hound (Egypt)	23½–27½	22½–26	50	45
Irish wolfhound (Ireland)	minimum 32, average 32–34	minimum 30	minimum 120	minimum 105
Norwegian elkhound (Norway)	20½	19½	55	48
Otterhound (England)	24–27	23–26	75–115	65–100
Petit Basset Griffon Vendéen (France)	13–15	same	25–35	same
Pharaoh hound (Egypt)	23–25	21–24	NA	NA
Rhodesian ridgeback (South Africa)	25–27	24–26	75	65
Saluki (Egypt)	23–28	may be considerably smaller	45–60	same
Scottish deerhound (Scotland)	30–32	28 or above	85–110	75–95
Whippet (England)	19–22	18–21	28	same
WORKING				
Akita (Japan)	26–28	24–26	75–110 or more	same
Alaskan Malamute (United States)	25	23	85	75
Bernese mountain dog (Switzerland)	25–27½	23–26	88	same
Boxer (Germany)	22½–25	21–23½	53–71	same
Bullmastiff (England)	25–27	24–26	110–130	100–120
Doberman pinscher (Germany)	26–28	24–26	66–88	same
Eskimo (Canada)†	20–27	same	60–105	same
Giant schnauzer (Germany)	25½–27½	23½–25½	66–78	same
Great Dane (Germany)	not less than 30, 32+ preferred	not less than 28, 30+ preferred	100+	same
Great Pyrenees (France)	27–32	25–29	100+	85+
Komondor (Hungary)	25½+	23½+	80–150	same
Kuvasz (Tibet)	28–30	26–28	100–115	70–90
Mastiff (England)	minimum 30	minimum 27½	175–190	same
Newfoundland (Canada)	28	26	130–150	100–120
Norwegian buhund (Norway)‡	17–18	same	26–40	same
Portuguese water dog (Portugal)	20–23	17–21	42–60	35–50
Rottweiler (Germany)	24–27	22–25	90–110	same
St. Bernard (Switzerland)	minimum 27½	minimum 25	110–200+	same
Samoyed (Siberia)	21–23½	19–21	50–65	same
Siberian husky (northeastern Asia)	21–23½	20–22	45–60	35–50
Standard schnauzer (Germany)	18½–19½	17½–18½	33	same

the bloodhound, best known for its remarkable ability to track. The Irish wolfhound, Scottish deerhound, basenji, whippet, and Norwegian elkhound are also in this group. In Canada, drevers belong to the Hound group as well, and in England the Grand Basset Griffon Vendéen is included.

Terriers. The Terrier group consists of both big and small dogs, but members of this group more than any other share a common ancestry and similar behavioral traits. Terriers were bred to rid barns and stables of vermin, to dig out unwanted burrowing rodents, and to make themselves generally useful around the stable. Terriers were used in the "poor man's recreation" of rat killing, especially in England where most of these breeds originated. Upper classes used terriers in foxhunting. They also were bred to fight each other in pits—hence the name pit bulls. During the late 1900s, dogfighting was outlawed in most states and countries of the Western world, and these dogs were thereafter bred for a friendly temperament rather than for aggressiveness.

Terrier characteristics

Terriers, because they had to fit in burrows and dig underground, were bred to stay relatively small, although large breeds are not uncommon. Their coats are usu-

Dog Breeds and Their Places of Origin (continued)

breed	height (inches)		weight (pounds)	
	dogs	bitches	dogs	bitches
TERRIERS				
Airedale terrier (England)	23	slightly less than 23	44	same
American Staffordshire terrier (England)	18–19	17–18	40–50	same
Australian terrier (Australia)	10–11	same	12–14	same
Bedlington terrier (England)	16½	15½	17–23	same
Border terrier (England)	11–12	same	13–15½	11½–14
Bull terrier (England)	21–22	same	52–62	same
Cairn terrier (Scotland)	10	9½	14	13
Dandie Dinmont terrier (England)	8–11	same	18–24	same
Fox terrier (wire and smooth; England)	maximum 15½	slightly smaller	18	16
Irish terrier (Ireland)	18	same	27	25
Kerry blue terrier (Ireland)	18–19½	17½–19	33–40	proportionately less
Lakeland terrier (England)	14½	13½	17	proportionately less
Manchester terrier (England)	14–16	same	7–22 (maximum)	same
Miniature bull terrier (England)	10–14	same	10–40 proportionate to height	same
Miniature schnauzer (Germany)	12–14	same	13–15	same
Norfolk terrier (England)	9–10	slightly smaller	11–12	same
Norwich terrier (England)	maximum 10	same	12	same
Scottish terrier (Scotland)	10	same	19–22	18–21
Sealyham terrier (Wales)	10½	same	23–24	slightly less
Skye terrier (Scotland)	10	9½	25	same
Soft-coated wheaten terrier (Ireland)	18–19	17–18	35–40	30–35
Staffordshire bull terrier (England)	14–16	same	28–38	24–34
Welsh terrier (Wales)	15–15½	proportionately smaller	20	same
West Highland white terrier (Scotland)	11	10	15–22	same
TOYS				
Affenpinscher (Germany)	9–11½	same	7–8	same
Brussels Griffon (Belgium)	7–8	same	8–10, maximum 12	same
Cavalier King Charles spaniel (England)†	12–13	same	10–18	same
Chihuahua (long- and smooth-coat; Mexico)	5	same	maximum 6	same
Chinese crested (hairless and powderpuff; China)	11–13	same	5–10	same
English toy spaniel (England)	10–10½	same	8–14	same
Italian greyhound (Greece or Turkey)	13–15	same	2 varieties: under 8/ over 8	same
Japanese Chin (Japan)	8–9	same	2 varieties: under 7/ over 7	same
Maltese (Malta)	5	same	4–6	same
Manchester terrier (toy; England)	6–7	same	7–12	same
Mexican hairless (Mexico)†	11–12	same	10 or less	same
Miniature pinscher (Germany)	10–12½	same	10	same
Papillon (Spain)	8–11	same	proportionate to height	same
Pekingese (China)	6–9	same	maximum 14	same
Pomeranian (Iceland, Lapland)	11 maximum	same	3–7	same
Poodle (toy; France)	maximum 10	same	7+	same
Pug (China)	10–11	same	14–18	same
Shih tzu (China)	8–11	same	9–16	same
Silky terrier (Australia)	9–10	same	8–10	same
Yorkshire terrier (England)	9	same	maximum 7	same
NON-SPORTING				
Bichon frise (Spain)	9½–11½	same	NA	NA
Boston terrier (United States)	15–17	same	15–25	same
Bulldog (England)	12–14	same	50	40
Chinese shar-pei (China)	18–20	same	40–55	same
Chow chow (China)	17–20	same	45–70	same
Dalmatian (Croatia)	19–23	same	50–55	same
Finnish spitz (Finland)	17½–20	15½–18	25–30	same
French bulldog (France)	12	same	maximum 28	same
Keeshond (Netherlands)	18	17	55–66	same
Lhasa apso (Tibet)	10–11	slightly smaller	13–15	same
Poodle (miniature; France)	10–15	same	NA	same
Poodle (standard; France)	minimum 15	same	45–70	same
Schipperke (Belgium)	11–13	10–12	18 maximum	same
Shiba inu (Japan)	14½–16½	13½–15½	20–30	same
Tibetan spaniel (Tibet)	10	same	9–15	same
Tibetan terrier (Tibet)	15–16	slightly smaller	20–24	same
HERDING				
Australian cattle dog (Australia)	18–20	17–19	35–45	same
Australian shepherd (United States)	20–23	18–21	35–70	same
Bearded collie (Scotland)	21–22	20–21	40–60	same
Belgian Malinois (Belgium)	24–26	22–24	62	same
Belgian sheepdog (Belgium)	24–26	22–24	62	same
Belgian Tervuren (Belgium)	24–26	22–24	62	same
Bouvier des Flandres (Belgium)	24½–27½	23½–26½	88	same
Briard (France)	23–27	22–25½	75	same
Cardigan Welsh corgi (Wales)	10½–12½	same	30–38	25–34
Collie (Scotland)	24–26	22–24	60–75	50–65
German shepherd dog (Germany)	24–26	22–24	75–95	same
Old English sheepdog (England)	minimum 22	minimum 21	66+	same
Pembroke Welsh corgi (Wales)	10–12	same	27	25
Puli (Hungary)	17	16	18–39	same
Shetland sheepdog (Scotland)	13–16	same	NA	same

*All breeds are recognized by the American Kennel Club unless otherwise noted. Height and weight measurements in boldface represent official standards of the AKC. All other measurements are average ranges for both dogs and bitches. Countries listed are those most closely associated with the breed. †Only recognized by the Canadian Kennel Club. ‡Only recognized by the Kennel Club of England.

ally rough and wiry for protection and require minimum maintenance. Unlike hounds or sporting dogs, which only found or chased their quarry, terriers were often required to make the actual kill as well, giving them a more pugnacious temperament than their size might suggest. They are usually lean with long heads, square jaws, and deep-set eyes. However, as with most breeds, form follows function: terriers that work underground have shorter legs, while terriers bred to work aboveground have squarer proportions. Terriers are active and vocal, naturally inclined to chase and confront.

The small terriers, which were often carried on horseback during foxhunts, were bred to be put to the ground. These dogs have very specific origins. In general, their names reflect the locale where the breed first took shape under the guidance of a small group of dedicated breeders. They are the Australian, Bedlington, border, cairn, Dandie Dinmont, Lakeland, Manchester, miniature schnauzer (of German origin), Norwich, Norfolk, Scottish, Sealyham, Skye, Welsh, and West Highland white. The larger terriers include the Airedale, Irish, Kerry blue, and soft-coated wheaten. In Canada, Lhasa apsos are part of this group. Britain claims the Parson Jack Russell and the Glen of Imaal terriers, both of which are found in the United States, although the Glen of Imaal is not recognized by the AKC.

Working dogs. This group of dogs was bred to serve humans in very practical and specific ways. They are the dogs most often associated with guarding, leading, guiding, protecting, pulling, or saving lives. Working dogs range in size from medium to large, but all are robust with sturdy and muscular builds. Working dogs are characterized by strength and alertness, intelligence and loyalty.

Among the breeds most often associated with guarding home, person, or property are the Akita, boxer, bullmastiff, Doberman pinscher, giant schnauzer, Great Dane, mastiff, Rottweiler, and standard schnauzer. Dogs bred to guard livestock are the Great Pyrenees, komondor, and kuvasz. In England, Pyrenean mountain dogs are recognized in this group, as are all the herding dogs, and, in Canada, Eskimo dogs are included. Also in the Working group are those dogs bred to pull, haul, and rescue. These include the Alaskan Malamute and Siberian husky, the Samoyed, the Bernese mountain dog, the Portuguese water dog, the Newfoundland, and the St. Bernard. Poodles of the three varieties (standard, miniature, and toy) are part of this group in England, as are several other breeds found in the Non-Sporting group in the United States.

Herding dogs. The Herding breeds are livestock-oriented, although they are versatile in protecting and serving humans in other ways. Herding breeds are intelligent and lively, making fine family pets or obedience competitors. Dogs were first used to assist sheepherders in the 1570s, but other varieties were bred for different herding tasks. Herding breeds are quick and agile, able to work on any terrain, and well-suited for short bursts of high speed. These dogs, even the compact breeds, are strong and muscular, possessing proud carriage of head and neck. Herding dogs perceive even the slightest hand signals and whistle commands to move a flock or seek out strays.

Some Herding breeds drive the flock by barking, or circling. Others, such as the Welsh Corgi, nip at the heels of the flock to drove them. The Border Collie confronts the flock with a silent stare, a form of intimidation which also proves effective.

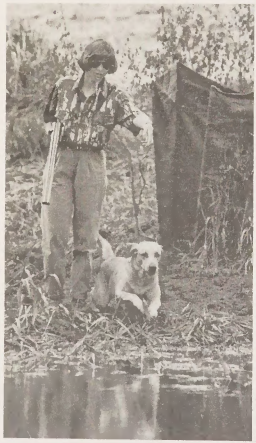
Uses for herding dogs

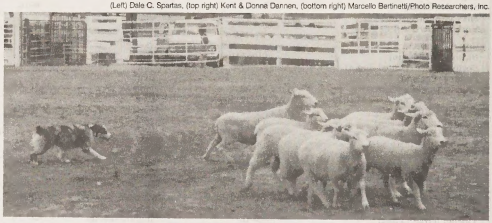
Herding dogs serve other functions. These breeds are excellent guards, used in the military and law enforcement, or for personal protection. Herding dogs are among those with the closest relationship to humans.

Toys. The Toy group is composed of those canines that were bred specifically to be companion animals. They were developed to be small, portable, and good-natured, the sort of dog that ladies of the court could carry with them. These dogs were largely pampered and treasured by aristocracy around the world. Several of these breeds come from ancient lineage. The Pekingese and the Japanese Chin were owned by royalty. No one else was permitted to own one of these breeds. They were carefully bred and nurtured, and until the mid-20th century they were not allowed to be exported out of their countries of origin. In England the cavalier King Charles spaniel, a bred-down version of a sporting spaniel, was the favourite pet of many royal families. Toy poodles also belong to this group.

The miniature pinscher resembles the Doberman pinscher but in fact is of quite different legacy. This perky little dog has a particularly distinctive gait, found in no

(Left) Dale C. Spartas, (top right) Kent & Donna Dannen, (bottom right) Marcello Bertinetti/Photo Researchers, Inc.







(Left) A Labrador retriever fetches game from the water; (top right) an Australian shepherd maneuvers a flock of sheep; (bottom right) Siberian huskies and Eskimo dogs pull an Inuit sled.

other breed. Its standard calls for a hackney gait, such as that found in carriage horses. Other members of the Toy group are equally individual in their looks and personalities, making this the most diverse group. They make ideal apartment or small-house pets and are found ranging from hairless (the Chinese crested) to the profusely coated Pekingese or Shih tzu. In general, however, Toy breeds are alert and vigorous dogs. They are fine-boned and well-balanced, often considered graceful animals.

Non-Sporting dogs. The Non-Sporting group is a catchall category for those breeds that do not strictly fit into any other group. (Arguments could be made for assigning some of these breeds to other groups. The Dalmatian, for instance, could be a Working dog, as it is in England.) This group includes the appealing bichon frise, the bulldog, the poodles (standard and miniature), and the Chinese shar-pei. All have unique histories, many quite ancient. Other Asian representatives are the Tibetan spaniel and the Tibetan terrier—neither of which are true spaniels or terriers—the chow chow, and the Lhasa apso. Non-Sporting is also the category for the Finnish spitz, the Keeshond, the French bulldog, and the schipperke. All the Non-Sporting breeds are of small to medium build with sturdy and balanced frames, often squarelike. The chow chow, French bulldog, and the Dalmatian are among the more muscular breeds in this group. In general, Non-Sporting dogs are alert and lively.

There is no comparable classification in Britain, although all these breeds, except for the Boston terrier, are found in other groups. The Boston terrier (not a true terrier although it once contained terrier blood) is one of the few native American dogs. (The others are the Alaskan Malamute, the beagle, the American foxhound, the Chesapeake Bay retriever, and the American cocker spaniel, all found in other groups.)

Breed standards. Purebred dogs are distinguished from mixed-breed animals because their genetic structure allows them to reproduce themselves generation after generation. Every breed that is registered with a national registry, such as the American Kennel Club or the Kennel Club of England, must have a "standard" for that breed. The standard is the blueprint by which a breed is evaluated. It describes the characteristics that make a particular breed unique. Standards were developed by fanciers who wanted to perpetuate a particular line or strain and who formed associations to foster certain breeds. It is the goal of most purebred-dog fanciers to breed dogs that best represent the ideal qualities for the breed as described by the standard. Standards outline requirements for physical traits and behavioral or "personality" traits.

Breed evaluation standards

RELATED CANIDS

The evolutionary process that brought about the domestication of the wild canid also created many other types of canids that have remained similar to dogs in genetic structure but with marked differences.

Wolves. The modern dog appears to be closely related to its most common ancestor, the wolf. The *Canis lupus* species includes more than 30 subspecies found in different parts of the world, some of which are now extinct. The subspecies vary greatly in size and colour, with the largest (averaging 95 to 100 pounds [43 to 45 kilograms]) found in the Arctic regions and the smallest (averaging 30 to 35 pounds) being the Texas red wolf.

The most striking similarities between the dog and the wolf are their instinctive behaviours of play, dominance and submission, scent marking, and the females' care for their young. Wolves are much more like dogs than like either coyotes or foxes in temperament and manners. Wolves appear to be instinctively more social than any of the other wild canids, thus lending themselves to interaction with humans in relationships beneficial to both. Wolves and dogs will mate willingly, as will dogs and coyotes. There are differences, however. The wolf matures more slowly than the dog. It reaches sexual maturity at about the age of two or three, at the same time that it achieves social maturity. A male wolf will not challenge the leaders of the pack until it is both physically and behaviorally mature. The female wolf cycles annually.

Coyotes. The coyote, *Canis latrans*, is a wide-ranging animal similar to wolves in some ways but different in others. Coyotes are light-boned, rangier in body with longer, narrower jaws and smaller ears and feet. They are thought to be the most intelligent of the wild canids because they have been able to survive and thrive despite human efforts to exterminate them for hundreds of years. Coyotes can weigh between 25 and 60 pounds and are usually gray to light tan in colour, depending on the region. There are more than a dozen subspecies of coyotes ranging throughout North and Central America. Coyotes tend to live in smaller groups than wolves, sometimes leading solitary lives until they reach sexual maturity at about two years, and they mate for life.

Foxes. Perhaps the most distinctive feature of the fox family, as compared with wolves and coyotes, is the eyes. They are yellow with elliptical pupils. All other canids, including dogs, have round pupils. Foxes are monogamous and do not live in packs. They are among the smaller species of canids, ranging from only 10 to 15 pounds. The more common foxes include the red fox, the gray fox, and the Arctic fox, which is valued for its fur.

There are several varieties of large-eared foxes, most of which are native to Africa and all of which are in danger of extinction because they are widely hunted and their habitat is being overbuilt by human settlements.

The smallest canid is the fennec. It weighs about three pounds, and its ears are about one-fourth of its body size. This endangered species is native to the desert areas of North Africa and the Arabian and Sinai peninsulas.

Jackals. There has been some disagreement over the years about whether the jackal is a true canid, but the four known varieties are now thought to be part of the same genus. Jackals are native from southeastern Europe into southern Asia, India, and Africa. The best-known variety is the golden jackal, which is a shimmery rust-gold in colour. Jackals are fleet-footed hunters, but they also eat insects and are best known as scavengers after larger animals, such as lions. The other varieties are the crafty black-backed jackal, the shy side-striped jackal, and the rare Abyssinian jackal.

Other wild canids. There are several different species of wild canids in South America. They fall somewhere between the fox and the wolf but are neither. One of the most interesting is the maned wolf, found in southern South America. The maned wolf is a fairly large animal, weighing about 50 pounds. It is very long-legged for its body length and is reddish brown in colour with a black ruff of hair around the neck. Its muzzle and feet are dark, and it has a white patch on the throat and a white plume on its tail.

In contrast to the maned wolf is the lowly Guiana bush dog. This short-legged, furry creature looks like a beaver but has longer legs and is an excellent swimmer and diver. It lives in packs of about 12 in the jungles of South America and is rarely seen by humans.

The raccoon dog

The most primitive member of the canid family is the Japanese raccoon dog. It is the only one that hibernates, moves into winter and summer ranges, and looks like a cross between a raccoon—because of its colour and markings—and a fuzzy fox. It has a heavy body (weighing a maximum of about 15 pounds) and is bred domestically for its fur, which is called tanuki.

One of the most important wild dogs is the dingo. It is believed that the dingo arrived in Australia between 9,000 and 15,000 years ago, but how it got there remains uncertain. Some scientists believe that the dingo is a small wolf, but others believe it is a true dog, much closer in behaviour to the domesticated dog than to the wolf. It has all the characteristics of a canine, with the exception that females cycle annually, like most of the other wild canids. Dingoes hunt in packs but may be found either alone or in a social group. Because dingoes are feared as livestock killers, considerable effort has been made to eliminate them, as with coyotes, although the latter have a higher survival rate. Dingoes are rarely seen in Australia now outside of zoos, and preservation efforts are being made to protect them in the wild.

Finally, the dhole, also called the Asiatic red dog, has

the widest range of any of the wild canids. It is found throughout most of the Asian mainland as high as the Himalayas and as low as the tropical islands of Borneo. It is reddish brown in colour, and on certain parts of the body the hair is gray or dark-tinged. The dhole is short-haired with a sturdy body and a pointed, felinelike face. Other varieties of the dhole family resemble short-coated wolves or Siberian-type dogs. Dholes hunt in packs; they do not bark or growl but may howl or whimper as a means of communication. Several of the canid species do not bark, but all are capable of sounding alarms or signaling to each other through vocalization.

Many of the wild dog species can be found only in captivity now. Through the efforts of zoologists, humans can maintain the link between these animals and the domestic dog that has thrived under human protection.

A distinction must be made between wild dogs and feral dogs. Feral dogs are domesticated dogs that have escaped to the wild, either through accident or neglect, and have reverted in the natural state to some of the characteristics inherent in all canids. They hunt or scavenge, run in packs, and become difficult to manage and train. They may become predators without the innate fear of humans that most wild canids have. Feral dogs may be found in cities or in the country and may be a reservoir of disease and a danger to domestic animals and people. (C.B.V.)

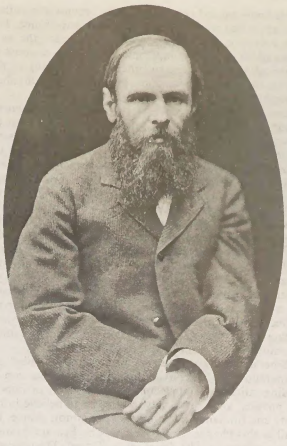
BIBLIOGRAPHY. THE AMERICAN KENNEL CLUB, *The Complete Dog Book*, 18th ed. (1992), comprehensively illustrates every breed that is registerable in the AKC stud book and includes a chapter on health care and puppy management. DAVID TAYLOR and CONNIE VANACORE, *The Ultimate Dog Book* (1990), is an illustrated overview of most of the breeds registered in the United States and Great Britain, with brief histories and descriptions of each.

FERNAND MÉRY, *The Life, History, and Magic of the Dog* (1970; originally published in French, 1968), traces the beginnings of the domestication of the dog and recites legends concerning the dog and its ancestors throughout the world. STANLEY J. OLSEN, *Origins of the Domestic Dog* (1985), is an anthropological study of fossils, primarily in the United States. MAXWELL RIDDLE, *The Wild Dogs in Life and Legend* (1979), describes many wild canids that exist in different parts of the world and relates the stories natives tell about them.

A vastly important contribution to the understanding of canine behaviour is JOHN PAUL SCOTT and JOHN L. FULLER, *Genetics and the Social Behavior of the Dog* (1965, reissued as *Dog Behavior: The Genetic Basis*, 1974), describing the genetic structure of personality based on original research by the authors. The foundational work describing the development of personality in puppies is CLARENCE J. PFAFFENBERGER, *The New Knowledge of Dog Behavior* (1963), based on research done by Scott and Fuller in the 1950s. JACK VOLHARD and MELISSA BARTLETT, *What All Good Dogs Should Know* (1991), is a basic primer for obedience training. CAROL LEA BENJAMIN, *Mother Knows Best: The Natural Way to Train Your Dog* (1985), uses a commonsense approach to teaching basic manners and solving problems.

WILLIAM J. KAY and ELIZABETH RANDOLPH, *The Complete Book of Dog Health* (1985), gives detailed descriptions of the major organ systems of the dog and describes common ailments and symptoms that dog owners can identify. HAROLD R. SPIRA, *Canine Terminology* (1982), definitively describes canine anatomy, illustrated from nose to tail. TERRI MCGINNIS, *The Well Dog Book* (1991), is a basic veterinary manual for dog owners. Also helpful is MALCOLM B. WILLIS, *Practical Genetics for Dog Breeders* (1992), which studies the genetic structure of the dog, including anatomy, coat colour, and breed differentiation. (C.B.V.)

Dostoyevsky



Dostoyevsky, 1880.
By courtesy of the Literary Museum of the Institute of Russian Literature, St. Petersburg

Fyodor Mikhaylovich Dostoyevsky, usually regarded as one of the finest novelists who ever lived, exercised an immense influence on 20th-century fiction and thought. Literary modernism, existentialism, and various schools of psychology, theology, and literary criticism have been profoundly shaped by Dostoyevsky's ideas. His works are often called prophetic because he so accurately predicted how Russia's revolutionaries would behave if they came to power. In his time he was also renowned for his activity as a journalist.

Major works and their characteristics. Dostoyevsky is best known for his novella *Notes from the Underground* and for four long novels, *Crime and Punishment, The Idiot, The Possessed* (also and more accurately known as *The Demons* and *The Devils*), and *The Brothers Karamazov*. Each of these works is famous for its psychological profundity, and, indeed, Dostoyevsky is commonly regarded as one of the greatest psychologists in the history of literature. He specialized in the analysis of pathological states of mind that lead to insanity, murder, and suicide and in the exploration of the emotions of humiliation, self-destruction, tyrannical domination, and murderous rage. These major works are also renowned as great "novels of ideas" that treat timeless and timely issues in philosophy and politics. Psychology and philosophy are closely linked in Dostoyevsky's portrayals of intellectuals, who "feel ideas" in the depths of their souls. Finally, these novels broke new ground with their experiments in literary form.

Family background

Early life. Dostoyevsky was born in Moscow on Oct. 30 (Nov. 11, New Style), 1821. His father, a retired military surgeon, served as a doctor at the Mariinsky Hospital for the Poor. Though a devoted parent, Dostoyevsky's father was a stern, suspicious, and rigid man. By contrast, his mother, a cultured woman from a merchant family, was kindly and indulgent. Dostoyevsky's lifelong attachment to religion began with the old-fashioned piety of his family, so different from the fashionable skepticism of the gentry.

Dostoyevsky's father bought an estate in 1831, and so young Fyodor spent the summer months in the country. Until 1833 Dostoyevsky was educated at home, before being sent to a day school and then to a boarding school. Dostoyevsky's mother died in 1837, and his father's death came suddenly in 1839. At the time, Dostoyevsky was a student in the Academy of Military Engineering in St. Petersburg, a career as a military engineer having been marked out for him by his father.

Dostoyevsky was evidently unsuited for such an occupation. Entranced with literature from an early age, he was drawn to Romantic and Gothic fiction, especially the works of Sir Walter Scott, Ann Radcliffe, Nikolay Karamzin, Friedrich Schiller, and Aleksandr Pushkin. Not long after completing his degree (1843) and becoming a sublieutenant, Dostoyevsky resigned his commission to commence a hazardous career as a writer living off his pen.

Early works. Dostoyevsky did not have to toil long in obscurity. No sooner had he written his first novella, *Bednyye lyudi* (1846; *Poor Folk*), than he was hailed as the great new talent of Russian literature by the most influential critic of his day, the "furious" Vissarion Belinsky.

Poor Folk

Poor Folk is cast in the then already anachronistic form of an epistolary novel. Makar Devushkin, a poor copying clerk who can afford to live only in a corner of a dirty kitchen, exchanges letters with a young and poor girl, Varvara Dobrosyolova. Her letters reveal that she has already been procured once for a wealthy and worthless man, whom, at the end of the novel, she agrees to marry. The novel is remarkable for its descriptions of the psychological (rather than just material) effects of poverty. Dostoyevsky transformed the techniques Nikolay Gogol used in "The Overcoat," the celebrated story of a poor copying clerk. Whereas Gogol's thoroughly comic hero utterly lacks self-awareness, Dostoyevsky's self-conscious hero suffers agonies of humiliation. In one famous scene, Devushkin reads Gogol's story and is offended by it.

In the next few years Dostoyevsky published a number of stories, including "Belyye nochi" ("White Nights"), which depicts the mentality of a dreamer, and a novella, *Dvoynik* (1846; *The Double*), a study in schizophrenia. The hero of this novella, Golyadkin, begets a double of himself, who mocks him and usurps his place. Dostoyevsky boldly narrates the story through one of the voices that sounds within Golyadkin's psyche so that the story reads as if it were a taunt addressed directly to its unfortunate hero.

Although Dostoyevsky was at first lionized, his excruciating shyness and touchy vanity provoked hostility among the members of Belinsky's circle. Nekrasov and Turgenev circulated a satiric poem in which the young writer was called, like Don Quixote, "The Knight of the Doleful Countenance"; years later, Dostoyevsky paid Turgenev back with a devastating parody of him in his novel *The Possessed.* Belinsky himself gradually became disappointed with Dostoyevsky's preference for psychology over social issues.

Political activity and arrest. In 1847 Dostoyevsky began to participate in the Petrashevsky Circle, a group of intellectuals who discussed utopian socialism. He eventually joined a related, secret group devoted to revolution and illegal propaganda. It appears that Dostoyevsky did not sympathize (as others did) with egalitarian communism and terrorism but was motivated by his strong disapproval of serfdom. On April 23, 1849, he and the other members of the Petrashevsky Circle were arrested. Dostoyevsky spent eight months in prison until, on December 22, the prisoners were led without warning to the Semyonovsky Square. There a sentence of death by firing squad was pronounced, last rites were offered, and three prisoners were led out to be shot first. At the last possible moment, the guns were lowered and a messenger arrived with the information that the tsar had deigned to spare their lives. The mock-execution ceremony was in fact part of the punishment.

Mock-execution ceremony

Dostoyevsky passed several minutes in the full conviction that he was about to die, and in his novels characters repeatedly imagine the state of mind of a man approaching execution. The hero of *The Idiot,* Prince Myshkin, offers several extended descriptions of this sort, which

readers knew carried special authority because the author of the novel had gone through the terrible experience. The mock execution led Dostoyevsky to appreciate the very process of life as an incomparable gift and, in contrast to the prevailing determinist and materialist thinking of the intelligentsia, to value freedom, integrity, and individual responsibility all the more strongly.

Instead of being executed, Dostoyevsky was sentenced to four years in a Siberian prison labour camp, to be followed by an indefinite term as a soldier. After his return to Russia 10 years later, he wrote a novel based on his prison camp experiences, *Zapiski iz myortvogo doma* (1861–62; *The House of the Dead*). Gone was the tinge of Romanticism present in his early fiction. The novel, which was to initiate the Russian tradition of prison camp literature, describes the horrors that Dostoyevsky actually witnessed: the brutality of the guards who enjoyed cruelty for its own sake, the evil of criminals who could enjoy murdering children, and the existence of decent souls amid filth and degradation. Tolstoy considered it Dostoyevsky's masterpiece. Above all, *The House of the Dead* illustrates that, more than anything else, it is the need for individual freedom that makes us human. This conviction was to bring Dostoyevsky into direct conflict with the radical determinists and socialists of the intelligentsia.

In Siberia Dostoyevsky experienced what he called the "regeneration" of his convictions. He rejected the condescending attitude of intellectuals, who wanted to impose their political ideas on society, and came to believe in the dignity and fundamental goodness of common people. He describes this change in his sketch "The Peasant Marey" (which appears in *The Diary of a Writer*). Dostoyevsky also became deeply attached to Russian Orthodoxy, as the religion of the common people, although his faith was always at war with his skepticism. In one famous letter he describes how he thirsts for faith "like parched grass" and concludes: "if someone proved to me that Christ is outside the truth, and that *in reality* the truth were outside of Christ, then I should prefer to remain with Christ rather than with the truth."

Epilepsy

Dostoyevsky suffered his first attacks of epilepsy while in prison. No less than his accounts of being led to execution, his descriptions of epileptic seizures (especially in *The Idiot*) reveal the heights and depths of the human soul. As Dostoyevsky and his hero Myshkin experience it, the moment just before an attack grants the sufferer a strong sensation of perfect harmony and of overcoming time. In 1857 Dostoyevsky married a consumptive widow, Mariya Dmitriyevna Isayeva (she died seven years later); the unhappy marriage began with her witnessing one of his seizures on their honeymoon.

Works of the 1860s. Upon his return to Russia, Dostoyevsky plunged into literary activity. With his brother Mikhail, he edited two influential journals, first *Vremya* ("Time"; 1861–63), which was closed by the government on account of an objectionable article, and then *Epokha* ("Epoch"; 1864–65), which collapsed after the death of Mikhail. After first trying to maintain a middle-of-the-road position, Dostoyevsky began to attack the radicals, who virtually defined the Russian intelligentsia. Dostoyevsky was repulsed by their materialism, their utilitarian morality, their reduction of art to propaganda, and, above all, their denial of individual freedom and responsibility. For the remainder of his life, he maintained a deep sense of the danger of radical ideas, and so his post-Siberian works came to be held in suspicion by the Soviet regime.

Notes from the Underground. In the first part of *Zapiski iz podpolya* (1864; *Notes from the Underground*) an unnamed first-person narrator delivers a brilliant attack on a set of beliefs shared by liberals and radicals: that it is possible to discover the laws of individual psychology, that human beings consequently have no free choice, that history is governed by laws, and that it is possible to design a utopian society based on the laws of society and human nature. Even if such a society could be built, the underground man argues, people would hate it just because it denied them caprice and defined them as utterly predictable. In the novella's second part the underground man recalls incidents from his past, which show him behaving, in answer to determinism, according to sheer spite. Dostoyevsky thus makes clear that the underground man's irrationalist solution is no better than the rationalists' systems. *Notes from the Underground* also parodied the bible of the radicals, Nikolay Chernyshevsky's utopian fiction *What Is to Be Done?* (1863).

Stay in western Europe. For several reasons, Dostoyevsky spent much of the 1860s in western Europe: he wanted to see the society that he both admired for its culture and deplored for its materialism, he was hoping to resume an affair with the minor author Appolinariya Suslova, he was escaping his creditors in Russia, and he was disastrously attracted to gambling. An unscrupulous publisher offered him a desperately needed advance on the condition that he deliver a novel by a certain date; the publisher was counting on the forfeit provisions, which would allow him nine years to publish all of Dostoyevsky's works for free. With less than a month remaining, Dostoyevsky hired a stenographer and dictated his novel *Igrok* (1866; *The Gambler*)—based on his relations with Suslova and the psychology of compulsive gambling—which he finished just on time. A few months later (1867) he married the stenographer, Anna Grigoryevna Snitkina. She at last put his life and finances in order and created stable conditions for his work and new family. They had four children, of whom two survived to adulthood.

The Gambler

Crime and Punishment. Written at the same time as *The Gambler, Prestupleniye i nakazaniye* (1866; *Crime and Punishment*) describes a young intellectual, Raskolnikov, willing to gamble on ideas. He decides to solve all his problems at a stroke by murdering an old pawnbroker woman. Contradictory motives and theories all draw him to the crime. Utilitarian morality suggests that killing her is a positive good because her money could be used to help many others. On the other hand, Raskolnikov reasons that belief in good and evil is itself sheer prejudice, a mere relic of religion, and that, morally speaking, there is no such thing as crime. Nevertheless, Raskolnikov, despite his denial of morality, sympathizes with the unfortunate and so wants to kill the pawnbroker just because she is an oppressor of the weak. His most famous theory justifying murder divides the world into extraordinary people, such as Solon, Caesar, and Napoleon, and ordinary people, who simply serve to propagate the species. Extraordinary people, he theorizes, must have "the right to transgress," or progress would be impossible. Nothing could be further from Dostoyevsky's own morality, based on the infinite worth of each human soul, than this Napoleonic theory, which Dostoyevsky viewed as the real content of the intelligentsia's belief in its superior wisdom.

Raskolnikov's Napoleonic theory

After committing the crime, Raskolnikov unaccountably finds himself gripped by "mystic terror" and a horrible sense of isolation. The detective Porfiry Petrovich, who guesses Raskolnikov's guilt but cannot prove it, plays psychological games with him until the murderer at last confesses. Meanwhile, Raskolnikov tries to discover the real motive for his crime but never arrives at a single answer. In a famous commentary, Tolstoy argued that there was no single motive but rather a series of "tiny, tiny alterations" of mood and mental habits. Dostoyevsky's brilliance in part lies in his complex rethinking of such concepts as motive and intention.

Crime and Punishment also offers remarkable psychological portraits of a drunkard, Marmeladov, and of a vicious amoralist haunted by hallucinations, Svidrigailov. Raskolnikov's friend Razumikhin voices the author's distaste for an ideological approach to life; Razumikhin's own life exemplifies how one can solve problems neither by grand ideas nor by dramatic gambles but by slow, steady, hard work.

Quite deliberately, Dostoyevsky made the heroine of the story, Sonya Marmeladova, an unrealistic symbol of pure Christian goodness. Having become a prostitute to support her family, she later persuades Raskolnikov to confess and then follows him to Siberia. In the novel's epilogue, the prisoner Raskolnikov, who has confessed not out of remorse but out of emotional stress, at first continues to maintain his amoral theories but at last is brought to true repentance by a revelatory dream and by Sonya's good-

ness. Critical opinion is divided over whether the epilogue is artistically successful.

The Idiot. Dostoyevsky's next major novel, *Idiot* (1868–69; *The Idiot*), represents his attempt to describe a perfectly good man in a way that is still psychologically convincing—seemingly an impossible artistic task. If he could succeed, Dostoyevsky believed, he would show that Christ-like goodness is indeed possible; and so the very writing of the work became an attempt at what might be called a novelistic proof of Christianity.

The work's hero, Prince Myshkin, is indeed perfectly generous and so innocent as to be regarded as an idiot; however, he is also gifted with profound psychological insight. Unfortunately, his very goodness seems to bring disaster to all he meets, even to the novel's heroine, Nastasya Filippovna, whom he wishes to save. With a remarkably complex psychology, she both accepts and bitterly defies the world's judgment of her as a fallen woman. Ippolit, a spiteful young man dying of consumption, offers brilliant meditations on art, on death, on the meaninglessness of dumb brutish nature, and on happiness, which, to him, is a matter of the very process of living.

Dostoyevsky's last decade. *The Possessed.* Dostoyevsky's next novel, *Besy* (1872; *The Possessed*), earned him the permanent hatred of the radicals. Often regarded as the most brilliant political novel ever written, it interweaves two plots. One concerns Nikolay Stavrogin, a man with a void at the centre of his being. In his younger years Stavrogin, in a futile quest for meaning, had embraced and cast off a string of ideologies, each of which has been adopted by different intellectuals mesmerized by Stavrogin's personality. Shatov has become a Slavophile who, like Dostoyevsky himself, believes in the "God-bearing" Russian people. Existentialist critics (especially Albert Camus) became fascinated with Kirillov, who adopts a series of contradictory philosophical justifications for suicide. Most famously, Kirillov argues that only an utterly gratuitous act of self-destruction can prove that a person is free because such an act cannot be explained by any kind of self-interest and therefore violates all psychological laws. By killing himself without reason, Kirillov hopes to become the "man-god" and so provide an example for human freedom in a world that has denied Christ (the God-man).

It is the novel's other plot that has earned Dostoyevsky the reputation of a political prophet. It describes a cell of revolutionary conspirators led by Pyotr Stepanovich Verkhovensky, who binds the group together by involving them in murdering Shatov. (This incident was based on the scheme of a real revolutionary of the time, Sergey Nechayev.) In lines that anticipate Soviet and Maoist cultural policy, Pyotr Stepanovich predicts that, when the revolution comes, "Cicero will have his tongue cut out, Copernicus will have his eyes put out, Shakespeare will be stoned," all in the name of "equality."

Pyotr is the son and Stavrogin the former student of the novel's weak but endearing liberal, Stepan Trofimovich Verkhovensky. Dostoyevsky suggests that the madness of the radical sons derives from their fathers' liberal skepticism, mockery of traditional morals, and, above all, neglect of the family.

A Writer's Diary and other works. In 1873 Dostoyevsky assumed the editorship of the conservative journal *Grazhdanin* ("The Citizen"), where he published an irregular column entitled "Dnevnik pisatelya" ("The Diary of a Writer"). He left *Grazhdanin* to write *Podrostok* (1875; *A Raw Youth,* also known as *The Adolescent*), a relatively unsuccessful and diffuse novel describing a young man's relations with his natural father.

A Raw Youth

In 1876–77 Dostoyevsky devoted his energies to *Dnevnik pisatelya,* which he was now able to bring out in the form he had originally intended. A one-man journal, for which Dostoyevsky served as editor, publisher, and sole contributor, the *Diary* represented an attempt to initiate a new literary genre. Issue by monthly issue, the *Diary* created complex thematic resonances among diverse kinds of material: short stories, plans for possible stories, autobiographical essays, sketches that seem to lie on the boundary between fiction and journalism, psychological analyses of sensational crimes, literary criticism, and political commentary. The *Diary* proved immensely popular and financially rewarding, but as an aesthetic experiment it was less successful, probably because Dostoyevsky, after a few intricate issues, seemed unable to maintain his complex design. Instead, he was drawn into expressing his political views, which, during these two years, became increasingly extreme. Specifically, Dostoyevsky came to believe that western Europe was about to collapse, after which Russia and the Russian Orthodox church would create the kingdom of God on earth and so fulfill the promise of the Book of Revelation. In a series of anti-Catholic articles, he equated the Roman Catholic church with the socialists because both are concerned with earthly rule and maintain (Dostoyevsky believed) an essentially materialist view of human nature. He reached his moral nadir with a number of anti-Semitic articles.

The *Diary's* most famous sections are usually known from anthologies and so are separated from the context in which they were designed to fit. These sections include four of his best short stories—"Krotkaya" ("The Meek One"), "Son smeshnogo cheloveka" ("The Dream of a Ridiculous Man"), "Malchik u Khrista na elke" ("The Heavenly Christmas Tree"), and "Bobok"—as well as a number of autobiographical and semifictional sketches, including "Muzhik Marey" ("The Peasant Marey"), "Stoletnaya" ("A Hundred-Year-Old Woman"), and a satire, "Spiritizm. Nechto o chertyakh Chrezychaynaya khitrost chertey, esli tolko eto cherti" ("Spiritualism. Something about Devils. The Extraordinary Cleverness of Devils, If Only These Are Devils").

The Brothers Karamazov. Dostoyevsky's last and probably greatest novel, *Bratya Karamazovy* (1879–80; *The Brothers Karamazov*), focuses on his favourite theological and philosophical themes: the origin of evil, the nature of freedom, and the craving for faith. A profligate and vicious father, Fyodor Pavlovich Karamazov, mocks everything noble and engages in unseemly buffoonery at every opportunity. When his sons were infants, he neglected them not out of malice but simply because he "forgot" them. The eldest, Dmitry, a passionate man capable of sincerely loving both "Sodom" and "the Madonna" at the same time, wrangles with his father over money and competes with him for the favours of a "demonic" woman, Grushenka. When the old man is murdered, circumstantial evidence leads to Dmitry's arrest for the crime, which actually has been committed by the fourth, and illegitimate, son, the malicious epileptic Smerdyakov.

The youngest legitimate son, Alyosha, is another of Dostoyevsky's attempts to create a realistic Christ figure. Following the wise monk Zosima, Alyosha tries to put Christian love into practice. The narrator proclaims him the work's real hero, but readers are usually most interested in the middle brother, the intellectual Ivan.

Alyosha

Like Raskolnikov, Ivan argues that, if there is no God and no immortality, then "all is permitted." And, even if all is not permitted, he tells Alyosha, one is responsible only for one's actions but not for one's wishes. Of course, the Sermon on the Mount says we are responsible for our wishes, and, when old Karamazov is murdered, Ivan, in spite of all his theories, comes to feel guilty for having desired his father's death. In tracing the dynamics of Ivan's guilt, Dostoyevsky in effect provides a psychological justification for Christian teaching. Evil happens not just because of a few criminals but because of a moral climate in which all people participate by harbouring evil wishes. Therefore, as Father Zosima teaches, "everyone is responsible for everyone and for everything."

The novel is most famous for three chapters that may be ranked among the greatest pages of Western literature. In "Rebellion," Ivan indicts God the Father for creating a world in which children suffer. Ivan has also written a "poem," "The Grand Inquisitor," which represents his response to God the Son. It tells the story of Christ's brief return to earth during the Spanish Inquisition. Recognizing him, the Inquisitor arrests him as "the worst of heretics" because, the Inquisitor explains, the church has rejected Christ. For Christ came to make people free, but, the Inquisitor insists, people do not want to be free, no matter what they say. They want security and certainty

"The Grand Inquisitor"

rather than free choice, which leads them to error and guilt. And so, to ensure happiness, the church has created a society based on "miracle, mystery, and authority." The Inquisitor is evidently meant to stand not only for medieval Roman Catholicism but also for contemporary socialism. "Rebellion" and "The Grand Inquisitor" contain what many have considered the strongest arguments ever formulated against God, which Dostoyevsky includes so that, in refuting them, he can truly defend Christianity. It is one of the greatest paradoxes of Dostoyevsky's work that his deeply Christian novel more than gives the Devil his due.

In the work's other most famous chapter, Ivan, now going mad, is visited by the Devil, who talks philosophy with him. Quite strikingly, this Devil is neither grand nor satanic but petty and vulgar, as if to symbolize the ordinariness and banality of evil. He also keeps up with all the latest beliefs of the intelligentsia on earth, which leads, in remarkably humorous passages, to the Devil's defense of materialism and agnosticism. The image of the "petty demon" has had immense influence on 20th-century thought and literature.

In 1880 Dostoyevsky delivered an electrifying speech about the poet Pushkin, which he published in a separate issue of *The Diary of a Writer* (August 1880). After finishing *Karamazov,* he resumed the monthly *Diary* but lived to publish only a single issue (January 1881) before dying of a hemorrhage on Jan. 28 (Feb. 9, New Style), 1881, in St. Petersburg.

MAJOR WORKS

NOVELS AND NOVELLAS: *Bednyye lyudi* (1846; *Poor Folk,* 1894); *Dvoynik* (1846; *The Double,* 1917); *Netochka Nezvanova* (1849; *Nyetochka Nyezvanov,* 1920); *Dyadyushkin Son* (1859; *Uncle's Dream,* 1888); *Selo Stepanchikovo i yego obitateli* (1859; *The Friend of the Family,* 1887); *Zapiski iz myortvogo doma* (1861–62; *Buried Alive; or, Ten Years of Penal Servitude in Siberia,* 1881, better known as *The House of the Dead*); *Unizhennyye i oskorblyonnyye* (1861; *Injury and Insult,* 1886, better known as *The Insulted and Injured*); *Zimniye zametki o letnikh vpechatleniyakh* (1863; *Winter Notes on Summer Impressions,* 1955); *Zapiski iz podpolya* (1864; *Letters from the Underworld,* 1913, better known as *Notes from the Underground*); *Prestupleniye i nakazaniye* (1866; *Crime and Punishment,* 1886); *Igrok* (1866; *The Gambler,* 1887); *Idiot* (1868–69; *The Idiot,* 1887); *Vechny muzh* (1870; *The Permanent Husband,* 1888, better known as *The Eternal Husband*); *Besy* (1872; *The Possessed,* 1913); *Dnevnik pisatelya* (1873–74, 1876–77, 1880, 1881; *The Diary of a Writer,* 1949); *Podrostok* (1875; *A Raw Youth,* 1916); *Bratya Karamazovy* (1879–80; *The Brothers Karamazov,* 1912).

SHORT STORIES: "Krotkaya" (1876; "The Gentle Maiden," 1913, also known as "The Meek One"); "Son smeshnogo cheloveka" (1877; "The Dream of a Queer Fellow," 1916, better known as "The Dream of a Ridiculous Man").

EDITIONS IN RUSSIAN AND IN ENGLISH TRANSLATION: The authoritative edition of Dostoyevsky's complete works is *Polnoe sobranie sochineniĭ F.M. Dostoevskogo,* 30 vol. (1972–90). As a rule, the best translations for capturing the artistic power of the fiction are still the ones in *The Novels of Fyodor Dostoevsky,* trans. by CONSTANCE GARNETT, 12 vol. (1912–20), available in many later editions, including in the *Modern Library* series. Since Garnett translated *The Possessed,* the missing chapter "At Tihon's," which his publisher had not allowed Dostoyevsky to publish, has come to light; the 1936 edition includes it in a rendition by Avrahm Yarmolinsky. Garnett's translation of *Notes from the Underground* was revised by RALPH E. MATLAW, *Notes from Underground, and The Grand Inquisitor* (1960), which also contains extracts from Chernyshevsky's work *What Is to Be Done?.*

Garnett has often been faulted for smoothing out the low and colloquial passages in Dostoyevsky. They are, however, admirably captured in *The Brothers Karamazov* (1990) and *Crime and Punishment* (1992), both trans. by RICHARD PEVEAR and LARISSA VOLOKHONSKY; unfortunately, these versions are not as successful as Garnett's in capturing other passages, especially lyrical prose. Another good rendition is *Notes from Underground,* trans. and ed. by MICHAEL R. KATZ (1989). *The Gambler, with Polina Suslova's Diary,* trans. by VICTOR TERRAS and ed. by EDWARD WASIOLEK (1972), is also of interest. *Netochka Nezvanova,* trans. by ANN DUNNIGAN (1970), is a fine version of this unfinished work.

For a long time, *The Diary of a Writer,* trans. by BORIS BRASOL, 2 vol. (1949, reprinted 1985), was the only, and extremely poor, translation. The superior version, *A Writer's Diary,* trans. by KENNETH LANTZ, 2 vol. (1993–94), is accompanied by helpful annotations explaining topical references. A selection of Dostoyevsky's journalism is *Dostoevsky's Occasional Writings,* trans. by DAVID MAGARSHACK (1963). His quasi-journalistic account of his first trip abroad is *Winter Notes on Summer Impressions,* trans. by DAVID PATTERSON (1988). Dostoyevsky's letters are available in *Complete Letters,* ed. and trans. by DAVID LOWE and RONALD MEYER, 5 vol. (1988–91).

The notebooks for Dostoyevsky's five long novels are available in editions that not only translate but also helpfully organize the Russian material; they are all edited by EDWARD WASIOLEK: *The Notebooks for Crime and Punishment* (1967), *The Notebooks for The Idiot* (1967), *The Notebooks for The Possessed* (1968), *The Notebooks for A Raw Youth* (1969), and *The Notebooks for The Brothers Karamazov* (1971). Additional notebooks are translated in *The Unpublished Dostoevsky: Diaries and Notebooks (1860–1881),* ed. by CARL R. PROFFER, 3 vol. (1973–76).

BIBLIOGRAPHY

Biography. The superior biography of Dostoyevsky (in any language) is the still incomplete multivolume study by JOSEPH FRANK, *Dostoevsky: The Seeds of Revolt, 1821–1849* (1976), *Dostoevsky: The Years of Ordeal, 1850–1859* (1983), *Dostoevsky: The Stir of Liberation, 1860–1865* (1986), and *Dostoevsky: The Miraculous Years, 1865–1871* (1995); these volumes also offer excellent portraits of the Russian intellectual milieu and illuminating readings of Dostoyevsky's works. Another good biography is LEONID GROSSMAN, *Dostoevsky* (1974; originally published in Russian, 2nd ed., 1965). One may also consult the diary of Dostoyevsky's mistress, Suslova, mentioned above; and the reminiscences of his second wife, ANNA DOSTOEVSKY, *Dostoevsky: Reminiscences* (1975; originally published in Russian, 2nd ed., 1971).

Criticism. Several studies survey Dostoyevsky's career with a chapter on each major work. The one to read first is KONSTANTIN MOCHULSKY, *Dostoevsky: His Life and Work* (1967; originally published in Russian, 1947). Others of note are EDWARD WASIOLEK, *Dostoevsky: The Major Fiction* (1964); MICHAEL HOLQUIST, *Dostoevsky and the Novel* (1977, reissued 1986); and RICHARD PEACE, *Dostoevsky: An Examination of the Major Novels* (1971, reissued 1992).

The most brilliant and most controversial book on Dostoyevsky is MIKHAIL BAKHTIN, *Problems of Dostoevsky's Poetics,* ed. and trans. by CARYL EMERSON (1984; originally published in Russian, 2nd ed., rev. and enlarged, 1963). Other classics of Russian criticism in English include a study by a prominent existentialist theologian, NICHOLAS BERDYAEV, *Dostoievsky,* trans. by DONALD ATTWATER (1934, reissued 1974; originally published in Russian, 1923). Also available is an essay originally published in Russian in 1903 by the Russian existentialist and Nietzschean LEV SHESTOV, "Dostoevsky and Nietzsche: The Philosophy of Tragedy," in his *Dostoevsky, Tolstoy, and Nietzsche* (1969), pp. 141–332.

DONALD FANGER, *Dostoevsky and Romantic Realism: A Study of Dostoevsky in Relation to Balzac, Dickens, and Gogol* (1965, reissued 1974), places Dostoyevsky in the context of European literature. On Dostoyevsky's obsessions and creative process, an outstanding, if diffuse, work is JACQUES CATTEAU, *Dostoyevsky and the Process of Literary Creation* (1989). The best study of Dostoyevsky's aesthetics is ROBERT LOUIS JACKSON, *Dostoevsky's Quest for Form: A Study of His Philosophy of Art,* 2nd ed. (1978), while his *The Art of Dostoevsky: Deliriums and Nocturnes* (1981), is especially good on *The House of the Dead.* ROBERT LOUIS JACKSON (ed.), *Dostoevsky: New Perspectives* (1984), is a fine critical anthology. Dostoyevsky's anti-Semitism is treated in DAVID I. GOLDSTEIN, *Dostoyevsky and the Jews,* trans. from French (1981); and in GARY SAUL MORSON, "Dostoevsky's Anti-Semitism and the Critics," *Slavic and East European Journal,* 27(8):302–317 (Fall 1983).

The outstanding study of *The Idiot* is ROBIN FEUER MILLER, *Dostoevsky and The Idiot: Author, Narrator, and Reader* (1981). On *Crime and Punishment,* ROBERT LOUIS JACKSON (ed.), *Twentieth Century Interpretations of Crime and Punishment* (1974), is a superb anthology of pithy extracts. Two divergent interpretations of *A Writer's Diary* by the same author are GARY SAUL MORSON, *The Boundaries of Genre: Dostoevsky's Diary of a Writer and the Traditions of Literary Utopia* (1981), and "Dostoevsky's Great Experiment," an introductory study to the Lantz translation of the *Diary* mentioned above. There are three outstanding books on *The Brothers Karamazov:* the reader should begin with ROBIN FEUER MILLER, *The Brothers Karamazov: Worlds of the Novel* (1992); and then turn to ROBERT L. BELKNAP, *The Structure of The Brothers Karamazov* (1967, reprinted 1989), and *The Genesis of The Brothers Karamazov: The Aesthetics, Ideology, and Psychology of Text Making* (1990).

(G.S.M.)

Drafting

Drafting provides graphic communication for the design and fabrication of machines, structures, systems, or various products. At the design stage, both freehand and mechanical drawings serve the functions of inspiring and guiding the designer and of communicating among the designer, collaborators, production department, and marketing or management personnel. Sometimes drafting is taken to mean only exact mechanical drawing procedures; at the design stage exact mechanical drawings can clarify, confirm, or disqualify a scheme that looked promising in a freehand sketch. Actually, both the sketch and the exact mechanical drawing are essential parts of the process of designing, and both belong to the field of drafting. After the basic design has been established, drafting skills aid in the development and transmission of the wealth of data necessary for the production and assembly of the parts. For an automobile, a skyscraper, or a spacecraft, tens of thousands of drawings may be needed to convey all of the requirements of the finished product from the designers to the fabricators.

The completion of the set of drawings necessary for the manufacture of a product or the construction of a project involves three important factors: (1) itemization of every detail and requirement of the final product or project; (2) application of good judgment and knowledge of standard drafting procedures to select the combination of drawings and specifications that will convey the information identified in stage (1) in the clearest possible manner; and (3) deployment of skilled personnel and suitable equipment to produce the documents specified in stage (2).

Drafting is based on the concept of orthographic projection, which in turn is the principal concern of the branch of mathematics called descriptive geometry. Although preceded by the publication of related material and followed by an extensive development, the book *Géométrie descriptive* (1798) by Gaspard Monge, an 18th-century French mathematician, is regarded as the first exposition of descriptive geometry and the formalization of orthographic projection. The growth and development of the drafting profession were favoured by the application of the concepts published by Monge, the need to manufacture interchangeable parts, the introduction of the blueprinting process, and the economy offered by a set of drawings that in most cases made the building of a working model unnecessary.

Persons with a variety of skills and specialties are essential to the design and implementation of engineering and architectural projects. Drafting provides communication among them and coordination of their activities. The designer has primary responsibility for the basic conception and final solution but depends upon the support of several levels of drafters who prepare graphic studies of details; determine fits, clearances, and manufacturing feasibility; and prepare the working drawings. The delineator, or technical illustrator, converts preliminary or final drawings into pictorial representations, usually perspective constructions in full colour to help others visualize the product, to inform the public, to attract investment, or to promote sales. Before undertaking their own drawings, persons entering the profession of drafting may trace drawings to revise or repair them, then advance to the preparation of detail drawings, tables of materials, schedules of subassemblies (such as doors and windows), and the dimensioning of drawings initiated by more experienced colleagues. The wide spectrum of activities demanded of a design team requires that its members combine experience and creativity with skills in visualization, analysis, and delineation and with knowledge of materials, fabrication processes, and standards.

It is the responsibility of the manufacturing, fabricating, or construction workers to follow a set of drawings and specifications exactly; there should be no need for them to ask questions or make decisions regarding particulars of the design. All such particulars are the responsibility of the design team; the drawings must clearly convey all necessary information so that the functional requirements of and regulatory restrictions on the completed product or project are satisfied, the mechanical properties of the materials are appropriate, and the machining operations and assembly or erection procedures are possible.

The strictly utilitarian objectives of drafting and its emphasis on clarity and accuracy clearly differentiate it from the allied art form covered in the article DRAWING. Cartographic drafting is treated in the article MAPPING AND SURVEYING. Some specific applications of drafting are dealt with in the articles BUILDING CONSTRUCTION: *Modern building practices;* DECORATIVE ARTS AND FURNISHINGS: *Interior design;* and INDUSTRIES, MANUFACTURING: *Clothing and footwear industry.*

This article is divided into the following sections:

Types of drawings 455
Dimensions and tolerances 455
Systems of representation 456
  Perspective
  Orthographic projection
  Descriptive geometry
  Ambiguity
  Hidden lines
  Auxiliary views
  Pictorial views
Drafting practice 459
  Standards
  Equipment
  Computers
Duplication of drawings 459
Bibliography 459

TYPES OF DRAWINGS

Varying according to the product or project, the set of drawings generally contains detail drawings (also called working drawings), assembly drawings, section drawings, plans (top views), and elevations (front views). For manufacturing a machine, the shape and size of each individual part, except standard fasteners, are described in a detail drawing, and at least one assembly drawing indicates how the parts fit together. To clarify interior details or the fitting together of parts, it may be necessary to prepare a section drawing, showing a part or assembly as though it had been cut by a plane, with a portion of the object removed. For constructing a building, plans, elevations, section drawings, and detail drawings are necessary to convey the information needed to estimate costs and then erect the structure. In this case the detail drawings contain exact information about such features as elevators, stairways, cabinetwork, and the framing of windows, doors, and spandrels. Different information appears in the set of drawings for a bridge, a dam, or a highway, but in each case the differences are related to the best manner of conveying the needed information.

DIMENSIONS AND TOLERANCES

The sizes of parts and overall sizes of assemblies are conveyed by dimensions placed on the drawing. The basic objective in dimensioning a drawing is to give the manufacturing or construction personnel the dimensions they need to do their work without requiring them to add, subtract, or estimate distances. If mass production is to be undertaken, special attention must be given to the dimensions of interchangeable parts that fit together. To dimension a distance as, say, two inches cannot require

that it be exactly two (2.000 . . .) inches, because no one can machine material with such precision. Particularly for parts of machinery, the designer must specify the acceptable range for the size of a hole, a shaft, or other feature requiring proper fit—perhaps 1.995 to 2.005 inches. The difference between the acceptable maximum and minimum dimensions given for a hole, shaft, or other feature is known as the tolerance. In the example above the tolerance is 0.010 (that is, 2.005 − 1.995) inch. Unsatisfactory tolerancy of mating parts ordinarily results in a machine with improper function or greatly reduced useful life. On the other hand, the cost of production increases greatly as tolerances are made stricter. It is an important design decision to require the correct level of tolerance for the functioning of any particular product. Additional information on a set of drawings indicates the materials to be used and the types of finish required on the surfaces.

SYSTEMS OF REPRESENTATION

Perspective. The shapes of all the parts and their interrelation are exactly described by the representation of that information in the set of drawings. Such description can be a lesser or greater challenge, depending on the complexity of the design. In the 15th century some of the leading artists and architects developed geometric schemes of perspective. Geometric perspective is a drawing method by which it is possible to depict a three-dimensional form as a two-dimensional image that closely resembles the scene as visualized by the human eye. The camera produces photographs with such resemblance. Images produced by the eye, the camera, and systems of perspective can all be interpreted in terms of what is known as central projection. Lines of sight may be thought of as extending from the points of the object under observation to a central point of convergence—the lens of the eye or the camera, or the reference point of the perspective construction. In the case of the eye these lines of sight are focused by the lens into an image on the curved retina. In the camera they pass through the lens to form an image on a flat piece of film. In systems of geometric perspective the converging lines of sight form an image on an imaginary picture plane located between the object and the central point of the construction.

The eye, the camera, and geometric perspective

Perspective drawings and photographs are easily interpreted because they closely resemble visual images. This resemblance includes the diminution of the relative size of the representations of portions of the object that recede from the viewer and the distortion of the angular relations of the lines of the object. The object shown in perspective in Figure 1A may be interpreted as a cube. The same object is represented in Figure 1B according to the projectional system ordinarily used for engineering and architectural drawings; there it is evident that the object is not a cube. Such projections are used because they convey accurate information about the shape of the object.

Orthographic projection. The projection used for engineering and architectural drawings is called orthogonal ("right-angled") or orthographic because the lines of sight from points on the object to the picture plane of the image are perpendicular to that plane. Thus, the lines of sight, called projectors, are parallel rather than convergent (as they are in the central projection of the eye, the camera, and geometric perspective).

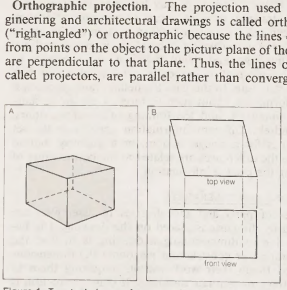


Figure 1: *Two techniques of representing an object.* (A) Perspective drawing, suggesting that the object is cubical. (B) Orthographic top and front views, revealing that the object is not cubical.

Descriptive geometry. Monge's reference system consisted of a vertical plane (*V* in Figure 2A) and a horizontal plane (*H*) that intersected in a ground line. As in Figure 2A, Monge numbered the four quadrants formed by *V* and *H* I, II, III, and IV. Figure 2A also shows two arrows, D_1 perpendicular to *H* and D_2 perpendicular to *V*. Each arrow represents the direction of projection from points on any object under study to one of the reference planes. Such an object is the L-shaped block located in the first quadrant. Monge introduced the concepts of the reference system, the formation of views by projectors perpendicular to the reference planes, the revolving of the *H* plane into coincidence with the *V* plane about the ground line as indicated by the curved arrows, and the retention of the images on the planes after the object had been removed and the *H* plane revolved. Figure 2B illustrates the final result: the projection on *V* is regarded as the front view, and the projection on *H* as the top view.

Monge's system

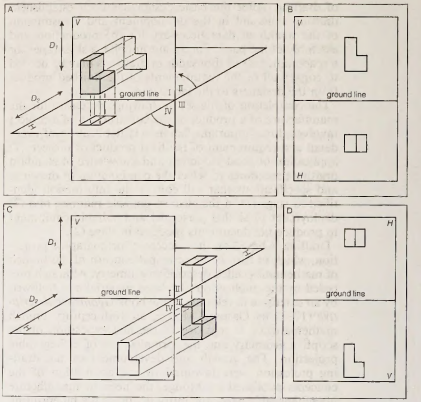


Figure 2: *Orthographic projections of a three-dimensional object onto vertical and horizontal planes.* (A) Object located in the first of four quadrants defined by intersecting planes. (B) Result of rotating the horizontal plane into coincidence with the vertical plane. (C) The same object located in the third quadrant. (D) Result of rotating the horizontal plane as in (B), showing exchanged positions of top and front views.

If the object were placed in the third quadrant (see Figures 2C and 2D), the projections would be exactly the same, but their relative locations on the paper would be reversed. If the object were located in the second quadrant, the two projections would have the same shape and size as in Figures 2B and 2D. Depending on the location of the object in the second quadrant, however, now either projection might be located above the other or one projection might overlap the other. The same is the case if the object were located in the fourth quadrant. This uncertainty is the reason that commercial use is limited to first- or third-quadrant projection. First-quadrant projection is often referred to as first-angle projection, and third-quadrant projection as third-angle projection.

Regardless of the quadrant (or angle) used, the views or projections are formed by the intersection of the projectors and the reference planes. Established conventions determine which points of the object are projected. If projectors were extended from every point on the object to the reference planes, the views would be silhouettes and would fail in their purpose of defining the object. The accepted rule is to project (1) all points on the edges between plane sur-

Choice of points to be projected

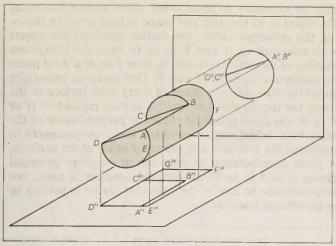


Figure 3: Selection of points to be projected in preparing orthographic views (see text).

faces that bound the object and (2) all points at which the projectors forming the view are tangent to curved surfaces of the object. Figure 3 illustrates these two sets of points. *AB* is the line of intersection of the cylindrical surface and plane surface *ABCD*. *CB* is the line of intersection of two plane surfaces. *EF* is not the line of intersection of two surfaces of the object, but projectors forming the top view are tangent to the cylindrical surface along the straight path from *E* to *F*, and thus E^HF^H properly appears in the top view. (The superscript *H* is used here to denote the projection on the *H* plane, and, similarly, *V* is used to denote the projection on the *V* plane.) *CD* is the line of intersection of the cylindrical surface and plane *ABCD*, but C^HG^H results from the tangency of projectors along the cylindrical surface. Every line projected in the identical front view of this object is a line of intersection of surfaces. *AD*, *BC*, and the plane *ABCD* all project as the same straight line in the front view because the plane *ABCD* is parallel to the projectors for that view.

Ambiguity. Ambiguity must be avoided in the views, dimensions, and notes of a set of drawings. Figure 4A shows pictorial representations of three different objects for which the identical front and top views in Figure 4B are correct. The ambiguity in the shape description provided by front and top views alone can be eliminated by adding a third, or side, view obtained by projecting the object onto a vertical plane perpendicular to *V*. In Figure 4B each set of three views describes only one of the objects without ambiguity.

In commercial or industrial practice, sets of drawings ordinarily provide at least three views of any part that is not a stamping, a gasket, a flat wrench, or other essentially two-dimensional form. Depending on the shape of the part, there may be a left-side view, a right-side view, or both. There may be reason for a back view, a bottom view, or both. Additional views are discussed below.

Hidden lines. It is standard practice to use dashes to represent any line of an object that is hidden from view.

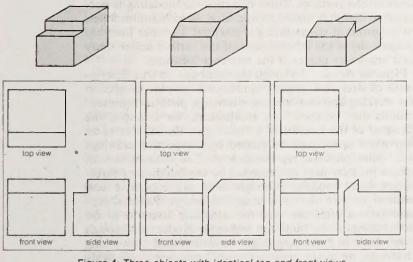


Figure 4: *Three objects with identical top and front views.* (Top row) Pictorial drawings. (Bottom row) Top, front, and side views, showing how the side views resolve the ambiguity.

A drafter—in deciding whether a line in a view should be represented as hidden or as visible—relies on the fact that in third-angle projection the near side of the object is near the adjacent view, but in first-angle projection the near side of the object is remote from the adjacent view. In Figure 4B (third-angle projection) the top of the front view is near the top view; the front of the top view is near the front view; and the front of the side view is near the front view. In first-angle projection, however, the top of the front view is remote from the top view; the front of the top view is remote from the front view, and the front of the side view is remote from the front view. In a third-angle projection, what is remote in an adjacent view cannot hide what is near in that view.

Advantages of third-angle projection

Figure 5 shows a pictorial representation of an object and the third-angle projections of that object. The arrangement of the three views gives intuitive reinforcement to the correct selection of the line shown as hidden in each view because it is blocked by portions of the object that are nearer in the adjacent views. The number of hidden lines in a view of a complicated object may be very great. For purposes of studying visibility, the direction of projection may be thought of as always vertically downward for a top view, always horizontally from front to back for a front view, and always horizontally right-to-left for a right-side view.

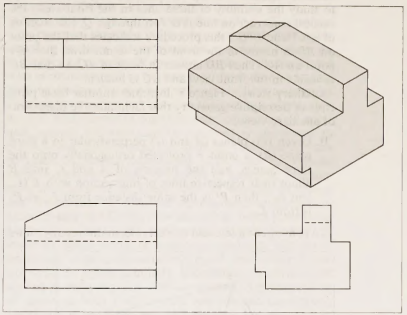


Figure 5: Use of dashed lines to represent edges hidden in views of a complicated object.

In Figure 5 the hidden lines in the views could be identified by visualizing the object, a process that can be quite difficult for complicated objects. The following basic principle of descriptive geometry is useful in analyzing such a problem:

I. If any point is projected orthogonally onto each of two perpendicular planes and the planes are rotated into coincidence about their line of intersection, then the projections of the point on the two planes will lie on a straight line perpendicular to the line of intersection.

Figure 6 demonstrates this statement. Although the ground line, or line of intersection of *H* and *V*, is seldom drawn in the representations of front and top or of front, top, and side views of objects, it is understood to be horizontal. Thus for any point *P*, P^H and P^V lie on a vertical line of the drawing.

A tetrahedron (triangular pyramid) with vertices *A*, *B*, *C*, and *D* is shown in third-angle projection in Figure 7. The edges *AC* and *BD* do not intersect, although their projections do. To determine which of these two edges is visible in the top view, the drafter considers location *M*, where the *H* projection of a point on *AC* and the *H* projection of a point on *BD* coincide. By principle I the *V* projections of these two points will lie on a vertical line from the crossing of A^HC^H and B^HD^H. A vertical construction line in Figure 7 indicates that the point on *BD*

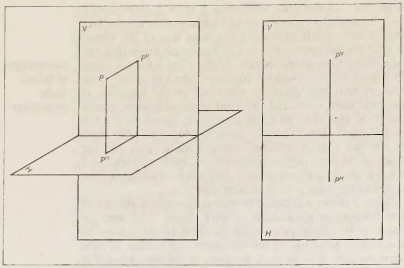


Figure 6: *Descriptive geometry, principle I (see text).* If a point is projected orthogonally onto two intersecting planes, which then are rotated into coincidence, the projections lie on a line perpendicular to the former line of intersection.

is nearer to the top of the tetrahedron than the point on *AC*. This means that *BD* crosses above *AC*, so that *BD* must be visible in the top view and *AC* hidden. Similarly, to study the visibility of these lines in the front view, the vertical construction line is drawn through *Q*, the crossing of A^1C^1 and B^1D^1; this procedure indicates that the point on *BD* is nearer to the front of the tetrahedron than the point on *AC*. Thus *BD* crosses in front of *AC*, so that *BD* is visible in the front view and *AC* is hidden.

Auxiliary views. Figure 8 illustrates another basic principle of descriptive geometry that facilitates the discussion of auxiliary views:

II. Given two planes (*A* and *C*) perpendicular to a third plane (*B*), a point *P* projected orthogonally onto the three planes, and the rotation of *A* and *C* into *B* about their respective lines of intersection with *B* (L_A and L_C), then P^A is the same distance from L_A as P^C is from L_C.

To convey complete and correct information many views may be necessary to show every plane surface bounding the object in its true size and shape at least once. In choosing the principal views, the drafter positions the object with reference to *H* and *V* so as to have the maximum number of its surfaces parallel to *H* or *V* or *R*, a third plane perpendicular to both *H* and *V*. Orthographic projection yields the true size and shape of every such surface in the front, the top, or the side view. A surface parallel to *H* or *V* or *R*, the three principal planes, is perpendicular to the other two. Additional or auxiliary views are necessary to represent the true size and shape of other plane surfaces. A plane perpendicular to only one of the three principal planes is said to be in an inclined position; a plane not perpendicular to any of the principal planes is said to be in an oblique position.

Inclined versus oblique planes

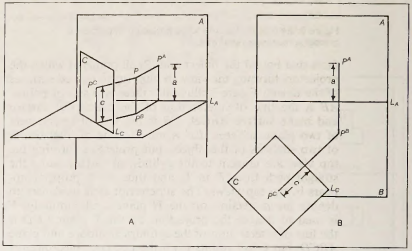


Figure 8: *Descriptive geometry, principle II (see text).* If a point is projected orthogonally onto three intersecting planes, of which two (*A* and *C*) are perpendicular to the third (*B*), and the three planes then are made to coincide, distances *a* and *c* must be equal. (A) Original arrangement of planes. (B) Planes rotated into coincidence.

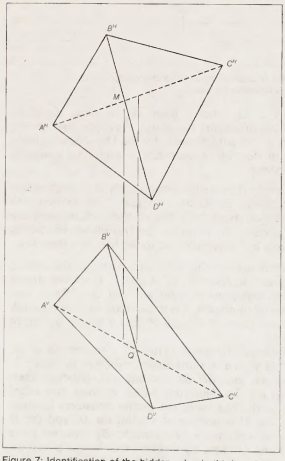


Figure 7: Identification of the hidden edge in third-angle projections of a tetrahedron (see text).

Figure 9 illustrates the application of principle II to represent the true size and shape of an inclined surface. The groove in surface *ABCD* makes an angle of 30° with a line (not shown) parallel to the edge *DC*. An auxiliary view in which *A*, *B*, *C*, and *D* are labeled with primes, obtained by projection onto a plane *P*, parallel to the surface *ABCD*, is the only one in which the true shape of *ABCD* and the true size of the 30° angle are correctly shown. The dimension indicated by the double-headed arrow is the same in the *H* (top) and auxiliary views, as required by principle II. The plane of the auxiliary view and the plane of the *H* view are perpendicular to the plane of the *V* view.

The true shape of an oblique surface can be shown correctly only on a second auxiliary view prepared by an extension of the procedure used for a first auxiliary view.

Automobile bodies, aircraft and ship hulls, and the irregular terrain of the natural site of a dam, bridge, or highway, are studied and detailed by means of contour lines on the surfaces. Three-dimensional modeling is necessary if design is highly competitive, as with automobiles, or if optimum streamlining is essential. Contour lines are projections of the intersections of the surface under study and imaginary planes at the reference locations.

Pictorial views. Although the emphasis on true descriptions of sizes and shapes requires orthographic projection for working and construction drawings, pictorial representations may be useful. In architecture, for example, the designer of the exterior of a building or the interior of an important space may be guided by perspective drawings and other pictorial representations. The construction of major projects may be preceded by the building of three-dimensional models, although these are expensive and seldom used in the early stages of design. Pictorial representations often are used for attracting investors or for advertising of new buildings and other products. Although a specialist in marketing might be intimidated by working drawings, he might grasp a pictorial representation easily enough to make useful suggestions about a design before production or construction was under way.

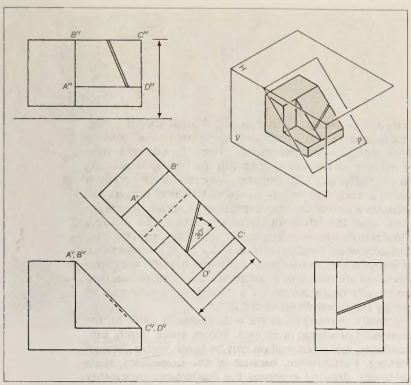


Figure 9: Use of auxiliary view to show true size and shape of an inclined surface (*ABCD*), which is not correctly represented in the front, top, or side view (see text).

The execution of a perspective drawing may require more time than is justified in the design of a small item. In many cases orthographic projection, coupled with the rotation of the object with respect to the reference planes, produces an adequate pictorial representation.

Oblique projection

Figures 2A, 2C, 3, 4A, and 5 illustrate the pictorial representation achieved by oblique projection, in which the principal surface of the object is considered to be in the plane of the paper and thus is represented in true size and shape. The angle the receding axis makes with the horizontal lines of the drawing is chosen arbitrarily but with care in terms of the clarity of the particular representation. True lengths are set off along the receding axis as an arbitrary choice. This is a convenient method for constructing a pictorial representation. Unacceptable distortion results when oblique projection is used to represent large objects or those with large dimensions or important details along the receding axis.

DRAFTING PRACTICE

Standards. The value of a set of drawings conveying the complete and correct information necessary for the execution of a project fostered the gradual standardization of practices. The widespread use of both first-angle and third-angle projection was long a major problem, but around the beginning of the 19th century a third-angle projection became the standard practice in the execution of industrial drawings in the United States. Australia followed this lead, but most industrial countries continued to follow Great Britain's use of first-angle projection. Architects in the United States and elsewhere generally use first-angle projection.

Drafting standards commonly evolve as a consensus develops among professional practitioners. Since 1917 in the United States the American National Standards Institute and its predecessors have encouraged this process and published standards for projections, various types of sections, dimensioning and tolerancing, representation of screw threads, all types of fasteners, graphic symbols for various specialties, and a great deal more. In other industrialized nations, analogous organizations—such as the British Standards Institution and the Deutsches Institut für Normung ("German Standards Institute")—function in the same way. In addition, many industrial groups and individual companies have established more detailed standards for their particular purposes.

The International Organization for Standardization, with headquarters in Geneva, coordinates global standards. International communication is hindered by the lack of agreement concerning first-angle versus third-angle projection and by the persistence in the United States of inches, feet, and other customary units for dimensioning. Economic pressures, however, are moving American industries to adopt the international metric system, SI units (Système Internationale d'Unités). The delay is related to the substantial costs of retooling and retraining. Because the strategy for correctly dimensioning a drawing is the same for all units, the rate of transition to SI units in the United States is not related to the drafting community, nor are SI units a special problem in drafting practice.

Equipment. Correct design information and projection are the imperatives of a set of engineering drawings. The skill and dexterity shown by some persons in drawing more accurately, more quickly, or more neatly have recognized value in the preparation of such drawings. Equipment has been invented to facilitate the performance of the manual tasks. Most widely known are the T square, triangle, protractor, and compass; the parallel straightedge is an alternative to the T square. The drafting machine, introduced about 1930, allows a straightedge to be moved while maintaining any desired angle between it and the edge of the drawing board. Combining the functions of the T square, triangle, protractor, and scale, it greatly increases the efficiency of producing a drawing.

Computers. A very important change in drafting procedure began in the early 1960s when programs were introduced to facilitate the composition of graphic images on the screen of a computer monitor, to retain the associated data in memory, and to retrieve the information to actuate plotting devices that produce not only the lines and arcs of an engineering drawing but also the symbols, dimension arrows, and strings of alphanumeric characters of notes and legends. Software can be prepared or purchased to perform the tasks involved in drafting: sketching of ideas to guide the design; calculation of the sizes of parts to satisfy codes, mechanical properties of materials, and machining requirements; preparation of working drawings; and production of pictorial representations. Computer-aided design (CAD) may be likened to word processing. Under direction, a word processor can correct misspellings, insert or delete words or sentences, rearrange sections of an article, or prepare accurately typed copies, but it cannot write an article. Similarly, knowledge, experience, and all but manual drawing skill are needed to produce a set of drawings with CAD, which has become increasingly important in industrial and architectural drafting.

DUPLICATION OF DRAWINGS

Blueprinting, the first economical method for duplicating drawings, was invented in 1842 and introduced in the United States in 1876. The diazo process, xerography, and computer-controlled drafting machines have more recently shared this function. The availability of numerous copies of drawings facilitated the division of labour among artisans, who formerly had worked out many details—such as exact sizes and shapes of parts, fits, and clearances—while custom building each item. The specification of these details became the duty of the designer-drafters, requiring them to refine their skills accordingly and leading to further development of the drafting profession.

BIBLIOGRAPHY. Aspects of drafting from basic instruction to industrial practices are treated in WALTER C. BROWN, *Drafting for Industry* (1974, reprinted 1984), a comprehensive treatment including coverage of CAD; PAUL WALLACH, *Metric Drafting* (1979), with emphasis on the use of the international metric system for dimensioning and tolerancing; and PAUL C. BARR *et al.*, *CAD: Principles and Applications* (1985), which covers general-purpose CAD functions and applications to industrial practice and training. WILLIAM T. GOODBAN and JACK J. HAYSLETT, *Architectural Drawing and Planning*, 3rd ed. (1979), discusses architectural sketching and drafting, including design concepts. See also GEORGE E. ROWBOTHAM (ed.), *Engineering and Industrial Graphics Handbook* (1982).

(R.A.K.)

Drawing

Drawing as formal artistic creation might be defined as the primarily linear rendition of objects in the visible world, as well as of concepts, thoughts, attitudes, emotions, and fantasies given visual form, of symbols and even of abstract forms. This definition, however, applies to all graphic arts and techniques that are characterized by an emphasis on form or shape rather than mass and colour, as in painting. Drawing as such differs from graphic printing processes in that a direct relationship exists between production and result. Drawing, in short, is the end product of a successive effort applied directly to the carrier, which is usually paper. Whereas a drawing may form the basis for reproduction or copying, it is nonetheless unique by its very nature.

Although not every artwork has been preceded by a drawing in the form of a preliminary sketch, drawing is in effect the basis of all visual arts. Often the drawing is absorbed by the completed work or destroyed in the course of completion. Thus, the usefulness of a ground plan drawing of a building that is to be erected decreases as the building goes up. Similarly, points and lines marked on a raw stone block represent auxiliary drawings for the sculpture that will be hewn out of the material. Essentially, every painting is built up of lines and pre-sketched in its main contours; only as the work proceeds is it consolidated into coloured surfaces. As shown by an increasing number of findings and investigations, drawings form the material basis of mural, panel, and book paintings. Such preliminary sketches may merely indicate the main contours or may predetermine the final execution down to exact details. They may also be mere probing sketches. Long before the appearance of actual small-scale drawing, this procedure was much used for monumental murals. With sinopia—the preliminary sketch found on a layer of its own on the wall underneath the fresco, or painting on freshly spread, moist plaster—one reaches the point at which a work that merely served as technical preparation becomes a formal drawing expressing an artistic intention.

Not until the late 14th century, however, did drawing come into its own—no longer necessarily subordinate, conceptually or materially, to another art form. Autonomous, or independent, drawings, as the name implies, are themselves the ultimate aim of an artistic effort; therefore, they are usually characterized by a pictorial structure and by precise execution down to details.

Formally, drawing offers the widest possible scope for the expression of artistic intentions. Bodies, space, depth, substantiality, and even motion can be made visible through drawing. Furthermore, because of the immediacy of its statement, drawing expresses the draftsman's personality spontaneously in the flow of the line; it is, in fact, the most personal of all artistic statements. It is thus plausible that the esteem in which drawing was held should have developed parallel to the value placed on individual artistic talent. Ever since the Renaissance, drawing has gradually been losing its anonymous and utilitarian status in the eyes of artists and the public, and its documents have been increasingly valued and collected.

This article will deal with the aesthetic characteristics, the mediums of expression, the subject matter, and the history of drawing.

This article is divided into the following sections:

General considerations 460
  Elements and principles of design 460
    Plane techniques
    The drawing surface
    Relationship between drawing and other art forms
  Surfaces, mediums, and techniques 462
    Types of ground
    Tools and techniques
  Applied drawings 470
  Subject matter of drawing 471
    Portraits
    Landscapes
    Figure compositions and still lifes
    Fanciful and nonrepresentational drawings
    Artistic architectural drawings
History of drawing 474
  Western 474
    14th, 15th, and 16th centuries
    17th, 18th, and 19th centuries
    Modern
  Eastern 476
Bibliography 477

General considerations

ELEMENTS AND PRINCIPLES OF DESIGN

The principal element of drawing is the line. Through practically the entire development of Western drawing, this figure, essentially abstract, not present in nature, and appearing only as a border setting of bodies, colours, or planes, has been the vehicle of a representational more or less illusionist rendition of objects. Only in very recent times has the line been conceived of as an autonomous element of form, independent of an object to be represented.

Conscious and purposeful drawing represents a considerable mental achievement, for the ability to reduce the spatial objects in the world around one to lines drawn on a plane presupposes a great gift for abstraction. The identification of the motif of a drawing by the viewer is no less an achievement, although it is mastered by practically all human beings. The visual interpretation of a line as a representation of a given object is made possible through certain forms of that line that call forth associations. The angular meeting of two lines, for example, may be considered as representing the borders of a plane; the addition of a third line can suggest the idea of a cubic body. Vaulting lines stand for arches, convergent lines for depth.

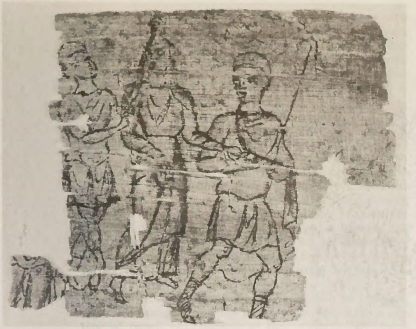
With the aid of this modest basic vocabulary, one can distill comprehensible images from a variety of linear phenomena. The simple outline sketch—Greek legend has it that the first "picture" originated from copying the shadows on the sand—represents one of the oldest and most popular possibilities of graphic rendition. After decisively characterizing the form of Egyptian drawing and the archaic art of Greece, the outline sketch became the chief vehicle of artistic communication in late antiquity and the Middle Ages. Used in a variety of ways in the early Renaissance, it became dominant once again in Neoclassicism, as it is, for that matter, in the classicist period of a given artist's total work.

The outline sketch

The outline sketch is elaborated into the detailed drawing by means of the line, which differentiates between the plastic and the spatial values of the object. Borders of individual objects, changes in the spatial plane, and varying intensities of colour applied within an outline sketch all tend to enrich and clarify the relationship between the whole and its component parts.

The free beginning, the disappearance, or the interruption of a line provides opportunities for gradually slurring an edge until it becomes a plane, for letting colour transitions fade away, for having the line vanish in the depth.

The thickening or thinning of a line can also be used to indicate, spatially or by means of colour, a change in



Late antique outline sketch, "The Abduction of Briseis," pen drawing on papyrus, Greek, 4th century AD. In the Bayerische Staatsbibliothek, Munich. 13.2 × 14.4 cm.

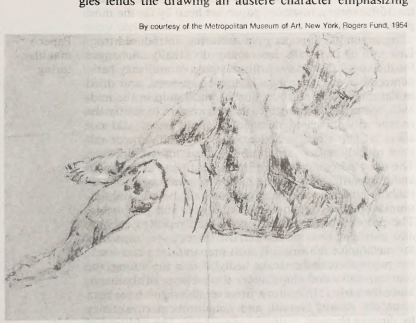
By courtesy of the Bayerische Staatsbibliothek, Munich

the object designated by that line. Even light-and-shadow values may be rendered by differences in stroke strength.

While the chopping up of a line into several brief segments, and, even more, the drawing of individual lines running parallel in one direction, makes the outlined form appear less corporeal and firm, it reproduces the visual impact of the form in a more pictorial manner. Slight shifts in the flow of the line are intended to represent smooth curves and transitions; they also reinforce the effect of light striking a surface and thus give the corporeal appearance. Finally, short, curving segments of a line that do not stand in a clearly angular relationship to one another but are arranged on the sheet in loose formation allow the pictorial and colour component to dominate, as in the work of the 16th-century Italian artist Jacopo Tintoretto. An extreme case is the complete dissolution of the linear stroke into dots and spots, as, for example, in the drawings of the 19th-century Pointillist painter Georges Seurat.

A mere combination of these varied shapes of the line, without reference to the mediums in which the lines are drawn, provides the artist with a plethora of subjective opportunities for the expression both of general stylistic traits and of personal characteristics. An arrangement of forceful, mainly straight strokes in accentuated, sharp angles lends the drawing an austere character emphasizing

By courtesy of the Metropolitan Museum of Art, New York, Rogers Fund, 1954



Loose use of line emphasizing pictorial effect, study after Michelangelo's "Day," by Jacopo Tintoretto (c. 1518–94), black and white chalk on blue paper. In the Metropolitan Museum of Art, New York. 34.9 × 50.5 cm.

dramatic and expressive traits. This method of drawing, in fact, is characteristic of stylistic epochs and artistic regions (not to mention individual artists) that prefer these qualities: in the rather sober city of Florence, in German Expressionism, where it is used to convey mood, but also in the drawings of Rembrandt and Vincent van Gogh. Soft lines, on the other hand, running in drawn-out, smoothly rounded forms and stressing graphic regularity above any statement of content, constitute the formal equivalent to elegant, courtly, and lyric qualities of expression. Accordingly, they are often found in drawings of the Soft style; in the early Renaissance, particularly in the work of artists from the Italian province of Umbria and in young Raphael's sketches; in the work of Nazarenes, a 19th-century group of Romantic painters whose subjects were mainly religious; in the Jugendstil, a late-19th- and early-20th-century German decorative style parallel to Art Nouveau in its organic foliate forms, sinuous lines, and non-geometric curves; and in a very pure form in one of the classic draftsmen, the 19th-century French painter Jean-Auguste-Dominique Ingres. A markedly even-stroke texture, with waxing and waning strokes in regular proportions and evenly distributed within the page, brings drawing close to calligraphic writing and is found in all stylistic epochs that value ornamentation.

Hatching and cross-hatching

The technique of hatching gives the line an additional potential for the clarification of plastic relationships and of light phenomena. In hatching, parallel, short, equidistant, more or less straight lines create static and tectonic (structural) values by marking individual body planes. Gently curved hatching stresses the roundness of the body and can also accentuate, as tone value, shaded parts of the representation.

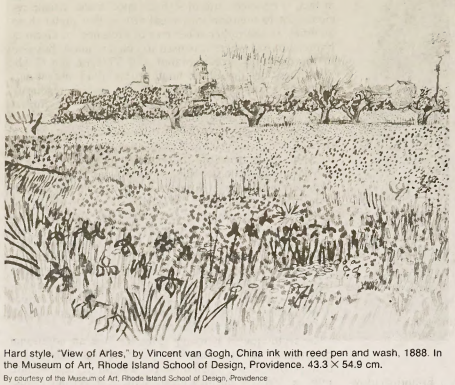
Cross-hatching, in which two layers of hatching intersect at right angles, reinforces the body-and-shadow effect. Known since the days of Michelangelo and Dürer in the 15th and 16th centuries, this artistic technique is often used with slanted or even curved hachures for the linear rendition of rounded parts. In rigorously monotone drawings, this method is the most suitable for the depiction of spherical bodies. The human body, with its highly articulated surface, can be modelled in this fashion very clearly and precisely. For 17th- and 18th-century engravers, this process became the most important means of drawing. All of these different possibilities of linear rendition can be achieved with pen and crayon as well as with the brush.

Plane techniques. Linear techniques of drawing are supplemented by plane methods, which can also be carried out with crayon. For example, evenly applied dotting, which is better done with soft mediums (see *Surfaces, mediums, and techniques* below), results in an areal effect in uniform tone. Various values of the chiaroscuro (pictorial representation in terms of light and shade without regard to colour) scale can also be rendered by means of dry or moist rubbing. Pulverized drawing materials that are rubbed into the drawing surface result in evenly toned areas that serve both as a closed foundation for linear drawing and as indication of colour values for individual sections.

More significant for plane phenomena, however, is brushwork, which, to be sure, can adopt all linear drawing methods but the particular strength of which lies in stroke width and tone intensity, a medium that allows for extensive differentiation in colour tone and value. Emphases created by the repeated application of the same tone provide illusionistic indentations that can be conceived of spatially and corporeally. Colour differences result from the use of various mediums. Brushwork also lends itself to spatial and plastic representation, just as it can constitute an autonomous value in nonrepresentational drawings.

All of these effects of monochrome drawing are accentuated with the use of varicoloured mediums of a basic material; for example, coloured chalks, drawing inks, or watercolour. While these mediums enrich the art of drawing, they do not widen its basic range.

The drawing surface. To these graphic elements must be added another phenomenon the formal significance of which is restricted to drawing: the effect of the unmarked drawing surface, usually paper. Almost all studies (draw-



Hard style, "View of Arles," by Vincent van Gogh, China ink with reed pen and wash. 1888. In the Museum of Art, Rhode Island School of Design, Providence. 43.3 × 54.9 cm.
By courtesy of the Museum of Art, Rhode Island School of Design, Providence

ings of details), many autonomous sheets, most portrait drawings, as well as figure compositions, still lifes, and even landscapes stand free on the sheet instead of being closed off with a frame-line. Thus, the empty surface, suggesting by itself a spatial background to the drawing on it, contributes actively to the artistic effect.

Details conveyed by empty space

Even within line composition, the surface left blank fulfills an essential role as a representational value defined by the drawn statements surrounding it as body of a given substantiality or space of a given expanse. Among the details conveyed by the empty space may be the planes of a face, the smooth width of a garment, the mass of a figure or object, the substance the borders and nuances of which are indicated by the drawing. Even the space around individual objects, the spatial distance between them and their environment, the width of a river and the depth of a landscape may be merely signalled by the drawing and filled by the void.

This void can itself become the dominant form enclosed by lines or contours—for example, in decorative sketches and in many ornamental drawings that make use of the negative form, an effect attainable also by tinting the blank planes.

Relationship between drawing and other art forms. The bond between drawing and other art forms is of course very close, because the preliminary sketch was for a long time the chief purpose of the drawing. A state of mutual dependence exists in particular between painting and drawing, above all, in the case of sketches and studies for the composition of a picture. The relationship is closest with preliminary sketches of the same size as the original, the so-called cartoons whose contours were pressed through or perforated for dyeing with charcoal dust. Once transferred to the painting surface, the sketch had served its purpose.

On autonomous sheets, too, the close connection between drawing and painting is evidenced by the stylistic features that are common to both. Drawing and painting agree in many details of content and form. Measurements; proportions of figures; relationship of figure to surrounding space; the distribution of the theme within the composition according to static order, symmetry, and equilibrium of the masses or according to dynamic contrasts, eccentric vanishing points, and overaccentuation of individual elements; rhythmic order in separate pictorial units in contrast to continuous flow of lines—all of these formal criteria apply to both art forms. The uniform stylistic character shared by drawing and painting is often less severely expressed in the former because of the spontaneous flow of the unfettered artist's stroke, or "handwriting," and of the struggle for form as recorded in the pentimenti (indications in the drawing that the artist had changed his mind and drawn over his original formulation). Furthermore drawing can stimulate certain aspects of movement more easily than painting can through the rhythmic repetition of a contour or the blended rubbing of a sharp borderline.

Still closer, perhaps, is the bond between drawing and engraving, which works with the same artistic means, with monochrome linearity as its main formal element and with various tone and plane methods closely related to those of drawing.

Drawing is more independent than sculpture because sculpture uses a three-dimensional model. As a result, sculptors' drawings can always claim a greater degree of autonomy. (For the special position of the architectural sketch, see *Subject matter of drawing* below.)

SURFACES, MEDIUMS, AND TECHNIQUES

Types of ground. One can draw on practically anything that has a plane surface (it does not have to be level); for example, papyrus and parchment, cloth, wood, metals, ceramics, and even walls, glass, and sand. (With some of these, to be sure, another dimension is introduced through indentations that give the visual effect of lines.) Ever since the 15th century, however, paper has been by far the most popular ground.

Paper manufacturing

The technique of paper manufacturing, introduced from East Asia by the Arabs, has remained virtually unchanged for the past 2,000 years. A fibrous pulp of mulberry bark, hemp, bast, and linen rags is drained, pressed, and dried in flat molds. The introduction of wood pulp in the mid-19th century, which enabled manufacturers to satisfy the enormously increased demand for bulk paper, did not affect art paper because paper of large wood content yellows quickly and is therefore ill-suited for art drawing. The essential preparation of the paper to give it a smooth and even surface for writing or drawing was once done by rubbing it with bone meal, gypsum chalk, or zinc and titanium white in a very thin solution of glue and gum arabic. The proper priming, achieved through repeated rubbing and polishing, was of the utmost importance, especially for metalpoint drawings. If such preparation is too weak, the paper accepts the stroke badly; if it is too strong, the coating cracks and chips under the pressure of the hand. Since the early 15th century, however, the sheets have been given the desired smooth and nonabsorbent consistency by dipping them in a glue or alum bath. The addition of glue also made it possible to impart to the pulp paper a quality that permitted pen drawings. Pigments, too, could of course be added to the pulp, and the so-called natu-

ral papers—chiefly blue and called Venetian papers after the centre of the retail trade in this commodity—became more and more popular. While the 17th century liked half tints of blue, gray, brown, and green, the 18th preferred warm colours such as ivory and beige, along with blue. Since the 18th century, paper has been manufactured in all conceivable colours and half tones.

The range of quality has also greatly increased since the end of the 18th century to give more painstakingly produced drawing papers. Even in earlier times, the absorbent Japan paper made of mulberry bark enjoyed great popularity. Handmade paper, stronger and free of wood, with an irregular edge, has remained to this day a favourite surface for drawings. Vellum, delicate and without veins, resembles parchment in its smooth surface. Modern watercolour paper is a pure linen paper glued in bulk and absolutely free of fat and alum; its two surfaces are of different grain. For pastel drawings, a firm, slightly rough surface is indicated, whereas pen drawings are best done on a very smooth paper.

Granulated and softer drawing tools, such as charcoal, chalk, and graphite are not as dependent on a particular type of paper; but, because of their slight adhesiveness, they often require a stronger bond with the foundation as well as some form of surface protection. This process of fixing was formerly done through repeated varnishing with gum-arabic solution and even with glue or egg-white emulsion. Modern siccatives (drying substances) inhibit discoloration but cannot prevent the living surface from appearing sealed, as it were, under a skin. In pastels especially, the manifold prismatic effects of finely powdered coloured crayons are thus lost, and the bright and airy surface is turned into an amorphous, heavy layer. Pastels, which brush off easily, are therefore best preserved under glass.

Tools and techniques. Such varied tools as slate pencils, charcoal, metal styli, and chalks may be used for drawing as well as all writing utensils, including pens, pencils, and brushes; indeed, even chisels and diamonds are used for drawing. Dry drawing tools differ in effectiveness from liquid ones because it is not irrelevant from the artistic point of view whether one uses a self-drawing medium that permits an evenly flowing line dependent only on hand pressure or a transferring tool that must be put down periodically and refilled, with resultant differences in the strength and concentration of the line. Modern drawing mediums that combine both possibilities, such as fountain pens, ball-point pens, and felt pencils, are recent inventions.

Dry and liquid drawing tools

No less varied than the nature and composition of these drawing mediums is their aesthetic effect. It would nevertheless be wrong to systematize the art of drawing on the basis of the techniques applied; not only does almost every technique have several applications but it can also be combined with other techniques, and the draftsman's temperament inevitably plays a role as well. Even if certain techniques predominate in certain periods, the selection of drawing mediums depends on the intended effect and not vice versa. Artists have always been able to attain the desired effect with a variety of techniques. Dry mediums, for example, are predestined for clear lines, liquid ones for plane application. Yet extremely fine strokes can also be made by brush, and broad fields can be marked in with pencil or crayon. Some mediums, including charcoal, one of the oldest, if not the oldest of all, allow both extremes.

Charcoal. In every hearth or fireplace, partially consumed pieces of wood remain that can be used as a convenient tool for drawing. Evidence of charcoal sketches for mural, panel, and even miniature paintings can still occasionally be seen under the pigment. Drawing charcoal produced from wood that is as homogeneous as possible gives a porous and not very adhesive stroke. The pointed charcoal pencil permits hair-thin lines; if used broadside on the surface, it creates evenly toned planes. Rubbing and pulverizing the charcoal line results in dimmed intermediate shades and delicate transitions. Because of its slight adhesiveness, charcoal is eminently suited to corrective sketching; but if the drawing is to be preserved, it must be protected by a fixative.

As a medium for quick, probing sketches and practice in studying models, charcoal was once much used in all academies and workshops. The rapid notation of difficult poses, such as Tintoretto demanded of his models, could be done quickly and easily with the adaptable charcoal pencil. While some of these sheets were deemed worthy of preservation, hundreds have surely been lost.

Charcoal used for portraits

Charcoal has often been used for portrait drawings to preserve for the eventual painting pictorial tints that were already present in the preliminary sketch. When destined to be autonomous portraits, charcoal drawings are executed in detail; with their sharp accents and delicate modelling, such portraits cover the whole range of the medium. In "Portrait of a Lady," by the 19th-century French painter Édouard Manet, the grain of the wood in the chair, the fur trimming on the dress, the compactness of the coiffure, and the softness of the flesh are all rendered in the same material: charcoal. Popular as that material was for studies and sketches, it has been used for independent drawings destined for preservation by only a few artists; for example, the 17th-century Dutch painter Paulus Potter. It is somewhat more frequent among the great draftsmen of the 19th and 20th centuries, such as Edgar Degas, Henri de Toulouse-Lautrec, Käthe Kollwitz, and Ernst Barlach.

Oiled charcoal, with the charcoal pencils dipped in linseed oil, provides better adhesion and a deeper black. Used in the 16th century by Tintoretto, this technique was applied above all by the Dutch draftsmen of the 17th century in order to set deep-black accents. The advantage of better adhesion in the indentations of the paper in contrast to dry charcoal, which sticks to the elevations, has to be paid for, however, by "incorrigibility"; *i.e.*, correction cannot be made. In addition, charcoal crayons that have been deeply dipped in oil show a brownish streak left by the oil alongside the lines.

Chalks. The chalks, which resemble charcoal pencils in outward appearance, are an equally important drawing medium. If charcoal was primarily a medium for quick sketching that could be corrected and for the search for artistic form, chalk drawing, which can also fulfill all of these functions, has steadily gained in importance as an autonomous vehicle of expression. Since the end of the 15th century, stone chalk, as found in nature, has become increasingly more significant in art drawing. As a basic material, alumina chalk has various degrees of hardness,

By courtesy of the National Museum Vincent van Gogh, Amsterdam



"Portrait of a Lady," charcoal drawing by Édouard Manet (1832–83). In the National Museum Vincent van Gogh, Amsterdam. 54 × 45 cm.

so that the stroke varies from slightly granular to homogeneously dense and smooth.

The attempt to produce a crayon or pencil of the greatest possible uniformity has led to the production of special chalks for drawing; that is, chalks, which, after being pulverized, washed, and molded into convenient sticks, allow a softer and more regular stroke and are also free of sandy particles. The admixture of pigments (carbons in the case of black chalks) creates various tints from a rich black to a brownish gray; compared to the much-used black chalk, the brown variety is of little significance. White chalk, also found in nature, is rarely employed as an independent medium for drawing, although it is frequently used in combination with other mediums in order to achieve reflections of light as individual accents of plastic modelling.

Beginning with the 15th century, chalk has been used increasingly for studies and sketches. Its suitability for drawing exact lines of any given width and also for laying on finely shaded tints makes it particularly appropriate for modelling studies. Accents that stress plastic phenomena are applied by varying the pressure of the hand. Characteristic details in portrait drawings in particular can be brought out in this manner. Pictorial values as well as light and shadow effects can be rendered with chalk without losing their firm, plastic form. For the same reason, chalk is also most valuable in sketching out paintings and indicating their values.

All of these qualities explain why chalk is such a good medium for autonomous drawings. Indeed, there is scarcely a draftsman who has not worked in chalk, often in combination with other mediums. Aside from portrait drawings done all over the world, landscapes have formed the main theme of chalk drawings, especially with the Dutch, in whose art landscape drawings have played a large role. Ever since the invention of artificial chalk made of lampblack (a fine, bulky, dull-black soot deposited in incomplete combustion of carbonaceous materials), which possesses a precisely measurable consistency—an invention ascribed to Leonardo da Vinci—the pictorial qualities of chalk drawing have been fully utilized. Chalks range from those that are dry and charcoal-like to the fatty ones used by lithographers.

Use of sanguine

Another very important drawing pencil is similarly a chalk product: the red pencil, or sanguine, which contains ferric oxide, which occurs in nature in shadings from dark brown to strong red and can also be manufactured from the same aluminum-oxide base with ferric oxide or rust added. Besides the stronger pictorial effect possible because of its chromatic value, sanguine also possesses a greater suppleness and solubility in water. Thus, a homogeneous plane can be created through moist rubbing, a compact stroke through liquid linear application, a very delicate tone through light wiping. Although this oxide was used for red tints in prehistoric painting, sanguine does not seem to have acquired artistic dignity until the 15th century, when it became customary to fix drawings by painting them over with a gum solution, for sanguine has no more adhesiveness than charcoal. In the 15th century, sanguine was a popular drawing medium because of its wealth of pictorial possibilities. Those inclined to be colorists—such as the portraitists Jean Clouet and Hans Holbein, the Flemish painters around Peter Paul Rubens, and, above all, the French artists of the 18th century—particularly favoured it. The possibilities of sanguine range from suggestive forms with markedly plastic values to a very pictorial, soft rendition of visual surface stimuli.

A combination of various chalks offers still richer coloristic possibilities. Black chalk and sanguine have been widely used since the 16th century to achieve colour differentiation between flesh tones, hair, and the material of garments. The combination of black and white chalk serves plastic modelling, as does that of the softer sanguine with white chalk; in the former case, the accentuation rests with the black, in the latter, with the more suggestive delineation in white.

A decidedly coloristic method lies in the combination of various chalk colours with one another and with tinted paper. Such pictorially executed sheets, called *à deux crayons* (with two colours) and *à trois crayons* (with three colours), respectively, were especially popular in 17th- and 18th-century France. Antoine Watteau reached a previously unheard of harmony of different chalks on natural paper. With the three colours, Nicolas Lancret, Jean-Étienne Liotard, Jacques-André Portail, François Boucher—to name but a few such artists—achieved sensitive drawings that are very appealing coloristically.

Pastel crayons

An additional colour refinement is made possible with pastel crayons. An ample selection of dry colour pigments in pastel crayons, prepared with a minimum of agglutinants and compounded with different shades of white for the articulation of tints, is commercially available. The colours can be laid on in linear technique directly with the crayons, but an area application made with a piece of soft suede or directly, with the fingers, is more frequent. Although this technique was known to the Accademia degli Incamminati (to the painter Guido Reni, for example) as early as the 17th century, it did not reach its flowering until the 18th century, especially in France (with Jean-Marc Nattier and Jean-Baptiste-Siméon Chardin) and in Venice (with Rosalba Carriera). Pastel chalks are particularly favoured for portraits; their effect approximates that of colour-and-area painting rather than line drawing.

In the 19th and 20th centuries, Degas reverted to a stronger accentuation of the delineatory aspects of drawing. With intermediate varnishes he achieved an overlay drawing with different colours and thus an increased emphasis on individual strokes. This technique, fundamentally different from the older one, was imitated with minor variations by Odilon Redon, Gustave Moreau, Jean-Édouard Vuillard, Pierre Bonnard, and others. It has also been borrowed by such Expressionist artists as Edvard Munch and Ernst Ludwig Kirchner.

Modern grease chalks offer a chromatic scale of similar range. Developed originally for such technical purposes as the lettering of very smooth surfaces, such as metal or glass, they can be applied in the same flat manner as pastels, although with the opposite aesthetic effect: that of compact colours. It was the 20th-century English sculptor Henry Moore who first and convincingly exploited the feasibility of continuing, with other mediums, such as pen or watercolour, work on the firm surface that had been led out with grease chalks.

Metalpoints. Metalpoints have been used for writing and delineation ever since the scriptoria of antiquity. It required little imagination to employ them also in drawing. The most frequently used material was soft lead, which on a smooth surface comes out pale gray, not very strong in colour, and easily erasable but very suitable for preliminary sketches. Aside from lead, tin and copper were also used, as well as sundry lead-and-pewter alloys. The 15th-century Venetian painter Jacopo Bellini's book of sketches in London with leadpoint drawings on tinted paper is a particularly valuable example of this technique, even if individual portions and, indeed, entire pages that had become effected were drawn over long ago. One can see little more than the traces left by the pencil because, as in many other metalpoint drawings, the sketches were redrawn in another medium. Botticelli, for example, sketched with a leadpoint the outline of his illustrations to Dante's *Divine Comedy*, retracing them afterward with the pen. Metalpoints were used into the 18th century for perspectivist constructions and auxiliary delineation, especially in architectural drawings.

Silverpoint drawing

More suited to permanent drawing is the silverpoint, which requires special preparation of the foundation and, once applied, cannot be corrected. Its stroke, also pale gray, oxidizes into brown and adheres unerasably. Silverpoint drawings accordingly require a clearer concept of form and a steady hand because corrections remain visible. Because too much pressure can bring about cracks in the foundation, the strokes must be even; emphases, modelling, and light phenomena must be rendered either by means of dense hachures, repetitions, and blanks or else supplemented by other mediums. Despite these difficulties, silverpoint was much used in the 15th and 16th centuries. Dürer's notebook on a journey to Holland shows landscapes, portraits, and various objects that interested him drawn in this demanding technique. Silverpoint was

much in demand for portrait drawings from the 15th into the 17th century; revived in the 18th-century Romantic era, it is still occasionally used by modern artists.

Graphite point. Toward the end of the 16th century, a new drawing medium was introduced and soon completely displaced metalpoint in sketching and preliminary drawing: the graphite point. Also called Spanish lead after its chief place of origin, this drawing medium was quickly and widely adopted; but because of its soft and smeary consistency it was used for autonomous drawings only by some Dutch painters, and even they employed it mostly in conjunction with other points. (It might be added that the graphite point was originally taken for a metal because its texture shines metallically in slanting light.) The lead pencil, or more properly *crayon Conté*, became established in art drawing after Nicolas-Jacques Conté invented, around 1790, a manufacturing process similar to that used in the production of artificial chalk. Purified and washed, graphite could henceforth be made with varying admixtures of clay and in any desired degree of hardness. The hard points, with their durable, clear, and thin stroke layers, were especially suited to the purposes of Neoclassicist and Romantic draftsmen. The Germans working in Rome, in particular, took advantage of the chance to sketch rapidly and to reproduce, in one and the same medium, subtle differentiations as well as clear proportions of plasticity and light. Among the most masterful pencil artists of all was Ingres, who presketched systematically in pencil the well-thought-out structure of his paintings.

The more pictorially inclined artists of the late 19th century, such as Ferdinand Delacroix, preferred softer pencils in order to throw into plastic relief certain areas within the drawing. Seurat, on the other hand, reached back to graphite in his drawings from the concert cafés, among them "Au concert européen" (Museum of Modern Art, New York) in which he translated the Pointillistic technique (applying dots of colour to a surface so that from a distance they blend together) into the monochrome element of drawing. Pencil frottage (rubbing made on paper laid over a rough surface), first executed by the Surrealist artist Max Ernst, represents a marginal kind of drawing, for here the artist's hand is no longer the sole creator of forms.

Pencil frottage

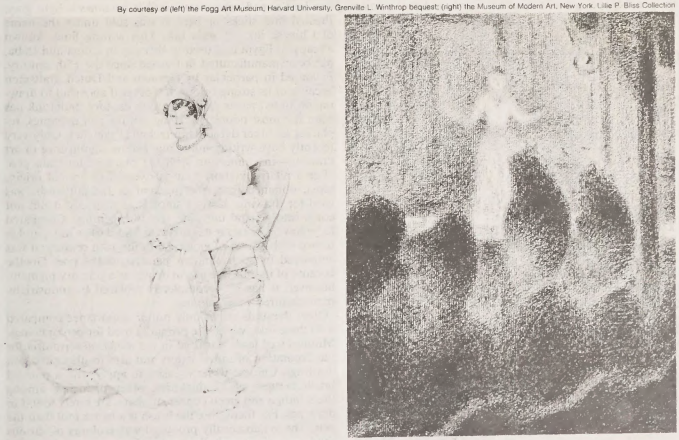
Coloured crayons. Coloured crayons, in circulation since the late 19th century, offer all the possibilities of black graphite points; and, in combinations, they attain a stronger colour value than chalks because they do not merge with one another. Every line preserves its original and characteristic colour, a form of independence that Gustav Klimt and Picasso exploited to the full.

Incised drawing. A role apart is that played by incised drawings. Their pronounced linearity gives them the visual appearance of other drawings; materially, however, they represent the opposite principle, that of subtracting from a surface rather than adding to it. Incised drawings are among the oldest documents of human activity. In primitive African cultures, the methods and forms of prehistoric bone and rock drawings have survived into the present. In a decorative and conceivably also symbolic form, incised decorations on pottery have existed for thousands of years; insofar as the comparison is valid, they correspond in every formal respect to applied drawings of the same period. A formal equivalent may also be observed in later times: in the decorative details of implements, especially metal—from the drawings on Greek mirrors, through the jewelry made at the end of the Roman Empire, to the scenes on medieval weapons and, above all, on Renaissance dress armour. More often than not these are drawings that follow certain models rather than free drawings in the sense of sketches.

Logically, one would also have to consider all niello work under the heading of drawing, because the picture in this case is cut out of the metal and filled with a deep black-coloured paste so that it appears to the eye as a linear projection on a plane. In like manner, work with the graver or burin (cutting tools) and with the etching needle on the engraving plate may be considered to parallel in its execution that gradual effort applied directly to the carrier that was defined earlier as the art of drawing. The difference lies in the fact that this work is not a goal in itself but the prerequisite for a printing process that is intended to be repetitive.

Brush, pen, and dyestuffs. Of the many possibilities of

By courtesy of (left) the Fogg Art Museum, Harvard University, Grenville L. Winthrop bequest; (right) the Museum of Modern Art, New York, Lillie P. Bliss Collection



Graphite mediums.
(Left) "Portrait of Mme. Guillaume Guillon Lethière," lead pencil drawing by Jean-Auguste-Dominique Ingres (1780–1867). In the Fogg Art Museum, Harvard University. 27 × 16.4 cm. (Right) "Au concert européen," conté crayon drawing by Georges Seurat, *c.* 1887. In the Museum of Modern Art, New York. 31.1 × 24.8 cm.

transferring liquid dyestuffs onto a plane, two have become particularly significant for art drawing: brush and pen. To be sure, finger painting, as found in prehistoric cave paintings, has occasionally been practiced since the late Renaissance and increasingly so in more recent times. For drawing as such, however, the method is irrelevant. Similarly, the use of pieces of fur, frayed pieces of wood, bundles of straw, and the like is more significant as a first step toward the camel's-hair brush than as indication that these objects were ever drawing mediums in their own right. Although it is antedated by the brush, which in some cultures (East Asia, for example) has remained in continued use, the pen has been the favorite writing and drawing tool ever since classical antiquity.

The principle of transferring dyestuffs with the pen has remained virtually unchanged for thousands of years. The capillary effect of the split tip, cut at a slant, applies the drawing fluid to the surface (parchment, papyrus, and, since the late Middle Ages, almost exclusively paper) in amounts varying with the saturation of the pen and the pressure exerted by the drawing hand. The oldest form

The reed pen

is that of the reed pen; cut from papyrus plants, sedge, or bamboo, it stores a reservoir of fluid in its hollow interior. Its stroke—characteristically powerful, hard, and occasionally forked as a result of stronger pressure being applied to the split tip—became a popular medium of artistic expression only with the rise of a subjective view of the artist's personality during the Renaissance. Rembrandt made superb use of the strong, plastic accents of the reed pen, supplementing it as a rule with other pens or brushes. Beginning in the 19th century with the Dutch artist van Gogh, pure reed-pen drawings with a certain forcefulness of expression have been created by many artists. Expressionists such as George Grosz used the reed pen frequently.

If the selection of the reed pen already implies a formal statement of sorts, that of the quill pen opens up a far wider range of possibilities. Ever since the rise of drawing in Western art—that is, since the late Middle Ages—the quill has been the most frequently used instrument for applying liquid mediums to the drawing surface. The importance accorded to this tool is attested by the detailed instructions in painters' manuals about the fashioning of the pen from wing shafts of geese, swans, and even ravens. The supple tip of the quill, available in varying strengths, permits a relatively wide scale of individual strokes—from soft, thin lines, such as those used in preliminary sketches for illustrations in illuminated books, through waxing and waning lines that allow differentiation within the stroke, to energetic, broad lines. It was only when metal pens began to be made of high-grade steel and in different strengths that they became a drawing implement able to satisfy the demands made by the individual artist's hand.

By courtesy of the Albertina, Vienna



"The Prophet Jonah Before the Walls of Nineveh," by Rembrandt, reed pen in bistre with wash, *c.* 1654–55. In the Albertina, Vienna. 21.7 × 17.3 cm.

Although all dyestuffs of low viscosity lend themselves to pen drawing, the various inks are most often employed. The manufacture of gallnut ink had been known from the medieval scriptoria (copying rooms set apart for scribes in monasteries). An extract of gallnuts mixed with iron vitriol and thickened with gum-arabic solution produces a writing fluid that comes from the pen black, with a strong hint of purple violet, and dries almost black. In the course of time it turns a darkish brown, so that the writing fluid in old manuscripts and drawings cannot always be identified by the colour alone. In contrast to other brown writing fluids, the more strongly coloured parts of gallnut ink remain markedly darker; and because inks of especially great vitriol content decompose the paper, the drawing, particularly in its more coloured portions, tends to shine through on the reverse side of the sheet. Only industrially produced chemical inks possess the necessary ion balance to forestall this undesirable effect.

Another ink, one that seems to have found no favour as a writing fluid but has nonetheless had a certain popularity in drawing, is bistre, an easily dissolved, light-to-dark-brown transparent pigment obtained from the soot of the lampblack that coats wood-burning chimneys. Its shade depends both on the concentration and on the kind of wood from which it is derived, hardwoods (especially oaks) producing a darker shade than conifers, such as pine. During the pictorially oriented Baroque period, in the 17th and early 18th centuries, the warm tone that can be thinned at will made bistre a popular medium with which to supplement the planes of a pen drawing.

Chinese or India ink

Also derived from a carbon base is India ink, made from the soot of exceptionally hard woods, such as olive or grape vines, or from the fatty lampblack of the oil flame, with gum-arabic mixed in as a binding agent. This deep-black, thick fluid preserves its dark tone for a long time and can be thinned with water until it becomes a light gray. Pressed into sticks or bars, it was sold under the name of Chinese ink or India ink. This writing fluid, known already in Egypt and used to this day in China and India, has been manufactured in Europe since the 15th century. Favoured in particular by German and Dutch draftsmen because of its strong colour, it lent itself above all to drawing on tinted paper. Since the 19th century, India ink has been the most popular drawing ink for pen drawings, replacing all other dyestuffs in technical sketches. Only very recently have writing inks gained some significance in art drawing—in connection with the practical fountain pen.

For a relatively short time, a dyestuff of animal origin, sepia, obtained from the pigment of the cuttlefish, was used for drawing. Known since Roman times, it did not come into general use until the 18th century. Compared to yellowish bistre, it has a cooler and darker tone, and is brown with a trace of violet. Until the 18th century, it was employed by such amateur painters as the poet Goethe because of its effectiveness in depth; as a primary pigment, however, it has been completely replaced by industrially manufactured watercolours.

Other dyestuffs are of only minor importance compared with these inks, which are primarily used for pen drawings. Minium (red lead) was used in the medieval scriptoria for the decoration of initial letters and also in illustrated pen drawings. Chinese white is easier to apply with a pointed brush because of its thickness; other pigments, among them indigo and green copper sulphate, are rarely found in drawings. For them, too, the brush is a better tool than the pen. The systematically produced watercolours of various shades are almost wholly restricted to technical drawings.

In combination with written texts, pen drawings are among the oldest artistic documents. Already in classical times, texts were illustrated with firm contours and sparse interior details. During the Middle Ages, marginal

drawings and book illustrations were time and again pre-sketched, if not definitively executed, with the pen. In book painting, decidedly delineatory styles developed in which the brush was also employed in the manner of a pen drawing: for example, in the Carolingian school of Reims, which produced the Utrecht Psalter in the 9th century, and also in southern Germany, where a separate illustrative form with line drawings was widespread with the *Biblia Pauperum* ("Poor People's Bibles," biblical picture books used to instruct large numbers of people in the Christian faith). The thin-lined outline sketch is also characteristic of the earliest individual drawings of the late Middle Ages and early Renaissance. Sketches after ancient sculptures or after nature as well as compositions dealing with familiar motifs form the main themes of these drawings. Such sheets were primarily used as models for paintings; gathered in sketchbooks, they were often handed on from one generation to the next. The practical usefulness of these drawings is attested by the supplements added to them by younger artists and by the fact that many metalpoint drawings that had become hard to decipher were redrawn with the pen, as shown by the sketchbooks of the 15th-century Italian artist Antonio Pisanello, now broken up and preserved in several different collections.

Use of sketch-books

In the 16th century, the artistic range of the pen drawing reached an individual articulation that it hardly ever attained again. Every artist was free to exploit with the pen the formal possibilities that corresponded to his talents. Thus Leonardo used a precise stroke for his scientific drawings; Raphael produced relaxed sketches, in which he probed for forms and variations of form; Michelangelo drew with short strokes reminiscent of chisel work; Titian contrasted light and dark by means of hachures laid broadly over the completed figures. Among the Northerners, Dürer mastered all the possibilities of pen drawing, from quick notation to the painstakingly executed autonomous drawing, ranging from a purely graphic and delineatory technique to a spatial and plastic modelling one; it is no wonder that he stimulated so many other artists. The subjective attitude of the later 16th century is often expressed more clearly in Mannerist drawings—characterized by spatial incongruity and excessive elongation of the human figures, which are as revelatory of the artist's personality as handwriting—than it is in completed works of painting and sculpture. A special form of exact drawing is found in models for engravings; some of these were directly mounted on the wood block; some anticipate the style of the copperplate engraving in the pen-drawing stage, with waxing and waning lines, delicate stroke layers, and cross-hatching for spatial and plastic effects.

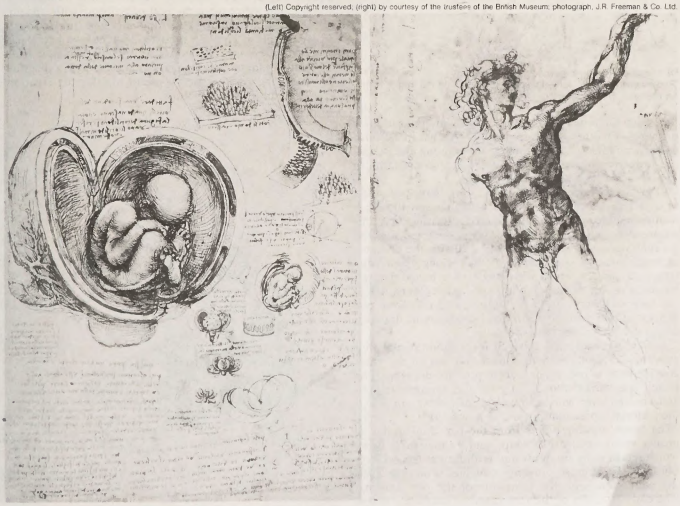
In the 17th century, the pen drawing took second place to combined techniques, especially wash, a sweep or splash of colour, applied with the brush. An open style of drawing that merely hints at contours, along with contrasting thin and powerful strokes, endowed the line itself with expressive qualities. In his numerous drawings, Rembrandt in particular achieved an exceedingly subtle plastic characterization and even light values through the differentiation of stroke layers and the combination of various pens and brushes.

Additional techniques came to the fore in the 18th century, with the pen sketch providing the scaffold for the drawing that was carried out in a pictorial style. Only decorative sketches and practical studies were laid out more often as linear drawings.

The closed, thin-contour drawing regained its importance with Neoclassicism at the end of the 18th century. The Nazarenes (the nickname of the Lucas Brotherhood—later Guild of St. Luke, who lived in monastic style) and Romantics consciously referred to the early Renaissance manner of drawing, modelling with thin lines. With closed contours, carefully set hair-and-shadow strokes, and precise parallel hachures, they attained plastic values by purely graphic means.

This technique was again followed by a more pictorially oriented phase, culminating in the late 19th century in

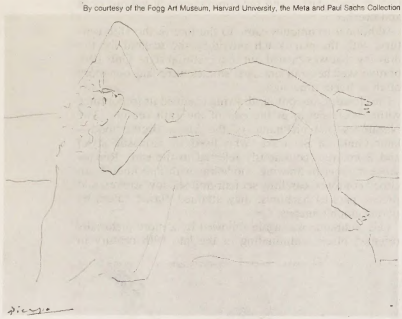
(Left) Copyright reserved, (right) by courtesy of the trustees of the British Museum; photograph, J.R. Freeman & Co. Ltd.



Individual expression in pen drawing of the Renaissance.
(Left) "Foetus in Utero," scientific drawing by Leonardo, pen and ink with red chalk, *c.* 1510–12. In the Royal Library, Windsor Castle. 30.1 × 21.4 cm. (Right) "Running Youth with Left Arm Extended," study for a sculpture by Michelangelo, pen and brown ink, *c.* 1504. In the British Museum. 37.5 × 22.9 cm.

the recognition of drawing as the most immediate and personal expression of the artist's hand. The pure pen drawing took its place by the side of other highly esteemed art forms. The English Art Nouveau artist Aubrey Beardsley at the end of the 19th century applied the direct black–white contrast to planes, while in the 20th century the French masters Henri Matisse and Picasso reduced the object to a mere line that makes no claim to corporeal illusion. A large number of illustrators, as well as the artists who draw the comic strips, prefer the clear pen stroke. In the Russian artist Wassily Kandinsky's nonrepresentational compositions, finally, the independence of the line as an autonomous formal value became a new theme in drawing. In the hair-thin automatist seismograms (so-called because of their resemblance to the records of earthquakes) of the 20th-century German artist Wols (Alfred Otto Wolfgang Schulze), which are sensitive to the slightest stirring of the hand, this theme leads to a new dimension transcending all traditional concepts of a representational art of drawing.

By courtesy of the Fogg Art Museum, Harvard University, the Meta and Paul Sachs Collection



Pure line pen and ink drawing of the 20th century. "Reclining Nude," by Pablo Picasso (1881–1973). In the Fogg Art Museum, Harvard University. 26 × 35 cm.

Although the brush is best suited to the flat application of pigments—in other words, to painting—its use in a clearly delineatory function, with the line dominating and (a crucial property of brush drawing) in monochrome fashion, can be traced back to prehistoric times.

All of the above-mentioned drawing inks have been used as dyes in brush drawings, often with one and the same pigment employed in combined pen-and-brush work. Still greater differentiation in tone is often obtained through concentrated or thinned mediums and with the addition of supplementary ones. To the latter belong chiefly distemper, a paint in which the pigments are mixed with an emulsion of egg or size or both, and watercolours, which can be used along with bistre and drawing ink. Even oils can sometimes be used for individual effects in drawing, as in the works of Jacob Jordaens.

Technique of sinopia

Sinopia, the preliminary sketch for a monumental wall painting, was done with the brush and has all the characteristics of a preparatory, form-probing drawing. The sketch was carried out directly on the appropriate spot and covered over with a thin layer of plaster, on which the pictorial representation was then painted.

The brush drawing differs from the pen drawing by its greater variation in stroke width, and by the stroke itself, which sets in more smoothly and is altogether less severely bordered. Early brush drawings nonetheless show a striking connection with the technique of the pen drawing. The early examples of the 15th century completely follow the flow of contemporaneous pen drawings. Leonardo's or Dürer's pen drawings, with their short, waxing and waning stroke layers, refine the system of pen drawing; many 16th-century artists used a comparable technique. The brush drawing for chiaroscuro sheets on tinted paper was popular because Chinese white, the main vehicle of delineation in this method, is more easily applied with the brush than the pen and because the intended pictorial effect is more easily attained, thanks to the possibility of changing abruptly to a plane representation.

Such representations are particularly distinctive as done by Vittore Carpaccio and Palma il Giovane in Venice and in a Mannerist spotting technique used by Parmigianino. In the 16th century, the brush nevertheless played a greater role as a supporting than as an independently form-giving instrument. Pure brush drawings were rare even in the 17th century, although the brush played a major role in landscapes, in which, by tinting of varying intensity, it ideally fulfilled the need to provide for all desired degrees of spatial depth and strength of lighting. Some Dutch artists, such as Adriaen Brouwer, Adriaen van Ostade, and Jan Steen, transcended the limits of drawing in the narrower meaning of the term by doing brushwork limited to a few tones within a monochrome scale, giving the impression of a pictorial watercolour.

Although the coloristically inclined 18th century was little interested in the restriction to a few shadings within one colour value, Jean-Honoré Fragonard raised this technique to perfection, with all its possibilities of sharply accented contours, soft delineation, delicate tones, and deep shadows. The brush drawings of the Spanish painter Francisco Goya must also be counted among the great achievements of this technique. In his strong plastic effects, the English painter George Romney made the most of the contrast between the white foundation and the broad brushstrokes tinted in varying intensities. Other English artists, among them Alexander Cozens, John Constable, and J.M.W. Turner, took advantage of the delicately graded pictorial possibilities for their landscape studies.

In the 19th century, the French artists Théodore Géricault, Eugène Delacroix, and Constantin Guys still followed the character of the brush drawing, even though it was already being replaced by the variegated watercolour and gouache painting, a method of painting with opaque colours that have been ground in water and mingled with a preparation of gum. In modern drawing, the brush has regained some importance as an effective medium for contrasting planes and as carrier of the theme; in this, the dry brush has proven itself a useful tool for the creation of a granular surface structure.

By courtesy of the Metropolitan Museum of Art, New York, Harris Brisbane Dick Fund, 1935



"Three Men Digging," by Francisco de Goya (1746–1828), brush drawing in sepia. In the Metropolitan Museum of Art, New York. 21 × 14 cm.

Pen and ink, chalk, silverpoint, and wash drawings



"Head of Girl" (study for "The Virgin of the Rocks," 1483), silverpoint on light brown paper by Leonardo da Vinci. In the Biblioteca Reale, Turin, Italy. 18.2 × 15.9 cm.



"An Island in the Lagoon," pen, brown ink, and carbon ink wash over ruled pencil lines by Canaletto (1697–1768). In the Ashmolean Museum, Oxford, England. 18.3 × 27.8 cm.



"Madonna with Many Animals," pen, ink, and watercolour by Albrecht Dürer, *c.* 1503. In the Albertina, Vienna. 32.1 × 24.3 cm.



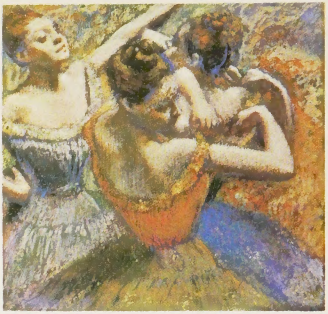
Study probably for "L'Indifferent," black, red, and white chalk on yellowish-gray paper by Jean-Antoine Watteau (1684–1721). In the Museum Boymans-van Beuningen, Rotterdam. 27.2 × 19 cm.

Plate 1: By courtesy of (top left) Biblioteca Reale, Turin, Italy, (top right) the Ashmolean Museum, Oxford, England, (bottom left) the Albertina, Vienna, (bottom right) the Museum Boymans-van Beuningen, Rotterdam



"Le Château Noir," pencil and watercolour by Paul Cézanne, *c.* 1895–1900. In the Museum Boymans-van Beuningen, Rotterdam. 36 × 52.6 cm.

Crayon, pencil, sepia, and pastel drawings



"The Dancers," pastel on paper by Edgar Degas, 1899. In the Toledo Museum of Art, Ohio. 62 × 64.5 cm.



"Trees and a Stretch of Water on the Stour," pencil and sepia wash by John Constable (1776–1837). In the Victoria and Albert Museum, London. 16.2 × 20.3 cm.



"Two Women III," crayon on paper by Willem de Kooning, 1952. In the Allen Memorial Art Museum, Oberlin College, Ohio. 37.5 × 47 cm.

Plate 2: By courtesy of (top) the Museum Boymans-van Beuningen, Rotterdam, (centre right) Victoria and Albert Museum, London, (bottom) Allen Memorial Art Museum, Oberlin College, Ohio; photograph, (centre left) The Toledo Museum of Art, Ohio, gift of Edward Drummond Libbey

The combination of various techniques plays a greater role in drawing than in all other art forms. Yet it is necessary, in the numerous drawings in which two or more mediums are involved, to distinguish between those in which the mediums were changed in the course of artistic genesis and those in which an artistic effect based on a combination of mediums was intended from the beginning.

In the first case, one is confronted with a preliminary sketch, as it were, of the eventual drawing: the basic structure with some variations is tried out in charcoal, chalk, metalpoint, pencil, or some other (preferably dry and easily corrected) material and then carried out in a stronger and more durable medium. Most pen drawings are thus superimposed on a preliminary sketch. The different materials actually represent two separate stages of the same artistic process.

More relevant artistically is the planned combination of different techniques that are meant to complement each other. The most significant combination from the stylistic point of view is that of pen and brush, with the pen delineating the contours that denote the object and the brush providing spatial and plastic as well as pictorial—that is, colour—values. The simplest combined form is manuscript illumination, where the delineated close contours are filled in with colour. The drawing may actually be improved if this is done by a hand other than the draftsman's or at a later time.

Pen-and-brush combination

More important is brushwork that supplements linear drawing, in which entire segments may be given over to one technique or the other; for example, the considerable use of white (which is hard to apply with the pen) in drawings on tinted paper. In similar complementary fashion the brush may be used for plastic modelling as a way of highlighting, that is indicating the spots that receive the greatest illumination. The technique of combined pen-and-brush drawing was favoured by the draftsmen of Germany and the Netherlands, especially in the circle around Dürer and the south German Danube School. Shadows, too, can be inserted in a drawing with dark paint. The illusion of depth can also be achieved with white and dark colours in a pure chalk technique.

In contrast to these methods, which still belong to a linear system of drawing, is the flat differentiation of individual segments of a work in (usually) the same medium: wash. Various bodies and objects are evenly tinted with the brush within or along the drawn contours. Planes are thus contrasted with lines, enhancing the illusionary effect of plasticity, space, and light and shadow. This modelling wash has been used again and again since the 16th century, sometimes in combination with charcoal, chalk, or pencil drawings. A further refinement, used particularly in landscape drawings, is wash in varying intensities; additional shadings in the sense of atmospheric phenomena, such as striking light and haze merging into fog and cloud, can be rendered through thinning of the colour or repeated covering over a particular spot. A chromatic element entered drawing with the introduction of diluted indigo, known in the Netherlands from the East India trade; it is not tied to objects but used in spatial and illusionist fashion, by Paul Brill and Hans Bol in the 16th and 17th centuries, for example. The mutual supplementation and correlation of pen and brush in the wash technique was developed most broadly and consistently in the 17th century, in which the scaffold, so to speak, of the pen drawing became lighter and more open, and brushwork integrated corporeal and spatial zones. The transition from one technique to the other—from wash pen drawings to brush drawings with pen accents—took place without a break. Claude Lorrain and Nicolas Poussin in 16th- and 17th-century France are major representatives of the latter technique, and Rembrandt once again utilized all its possibilities to the full.

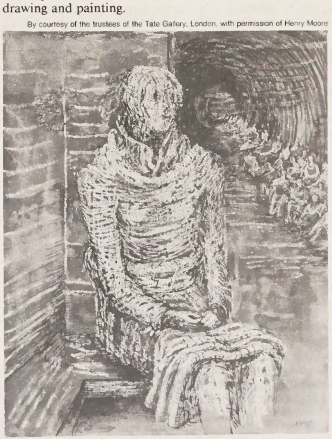
Whereas this method served—within the general stylistic intentions of the 17th century—primarily to elucidate spatial and corporeal proportions, the artists of the 18th century employed it to probe this situation visually with the aid of light. The unmarked area, the spot left empty, has as much representational meaning as the pen contours, the lighter or darker brush accent, and the tinted area.

The art of omission plays a still greater role, if possible, in the later 19th century and in the 20th. Paul Cézanne's late sheets, with their sparse use of the pencil and the carefully measured out colour nuances, may be considered the epitome of this technique. As the colouring becomes increasingly varied through the use of watercolours to supplement a pen or metalpoint drawing, one leaves the concept of drawing in the strict sense of the term. According to the quality and quantity of the mediums employed, one then speaks of "drawings with watercolour," "watercolourized drawings," and "watercolours on preliminary drawings." The predominant stroke character, rather than the fact that paper is the carrier, is the chief feature when deciding whether or not the work may legitimately be called a drawing.

Combination of dry and fluid mediums

The combination of dry and fluid drawing mediums provides a genuine surface contrast that may be exploited for sensuous differentiation. Here again a distinction must be made between various ways of applying the identical medium—for example, charcoal and charcoal dust in a water solution or, more frequently, sanguine and sanguine rubbed in with a wet brush—and the stronger contrast brought about by the use of altogether different mediums. Chalk drawings are frequently washed with bistre or watercolour, after the principle of the washed pen drawing. Stronger contrasts, however, can be obtained if the differing techniques are employed graphically, as the Flemish draftsmen of the 17th century liked to do. The Chinese ink wash of chalk drawings also contributes to the illusion of spatial depth. Along with such Dutch painters as Jan van Goyen and members of the family van de Velde, Claude Lorrain achieved great mastery in this technique. The differentiated treatment of the foreground with pen and brush and the background with chalk renders spatial depth plausible and plastic. In modern art, the use of different mediums—whether for plastic differentiation, such as Henry Moore carried out with unequalled mastery in his "Shelter Drawings," or only for the purpose of contrasting varied surface stimuli of nonrepresentational compositions as well as the enrichment with colours and even with collage elements (the addition of paper, metal, or other actual objects) broadens the concept of the drawing so that it becomes an autonomous picture the mixed technique of which transcends the borderline between drawing and painting.

By courtesy of the trustees of the Tate Gallery, London, with permission of Henry Moore



Use of different mediums to emphasize sculptural effects. "Women Seated in the Underground," crayon and wash drawing by Henry Moore, 1941. In the Tate Gallery, London. 48.3 × 38.1 cm.

Mechanical devices. Mechanical aids are far less important for art drawing than for any other art form. Many draftsmen reject them altogether as unartistic and inimical to the creative aspect of drawing.

Apart from the crucial importance that mechanical aids have had and continue to have for all kinds of construction diagrams, plans, and other applied drawings, some mechanical aids have been used in varying but significant measure for artistic drawings. The ruler, triangle, and compass as basic geometric instruments have played a major role, especially in periods in which artists created in a consciously constructionist and perspectivist manner. Marks for perspective constructions may be seen in many drawings of early and High Renaissance vintage.

For perspectively correct rendition, the graticulate frame, marked off in squares to facilitate proportionate enlargement or reduction, allowed the object to be drawn to be viewed in line with a screen on the drawing surface. Fixed points can be marked with relative ease on the resultant system of coordinates. For portrait drawings, the glass board used into the 19th century had contours and important interior reference points marked on it with grease crayons or soap sticks, so that they could be transferred onto paper by tracing or direct copying. Both processes are frequently used for preliminary sketches for engravings to be duplicated, as is the screened transmission of a preliminary sketch onto the engraving plate or, magnifying, the painting surface. In such cases the screen lies over the preparatory drawing.

Mirrors and mirror arrangements with reducing convex mirrors or concave lenses were likewise used (especially in the 17th and 18th centuries) as drawing aids in the preparation of reproductions. Even when it was a matter of the most exact rendition of topographical views, such apparatus, as well as the camera obscura (a darkened enclosure having an aperture usually provided with a lens through which light from external objects enters to form an image on the opposite surface), were frequently employed. In a darkened room the desired section is reflected through a lens onto a slanting mirror and from that inverse image is reflected again onto the horizontally positioned drawing surface. Lateral correction can be obtained by means of a second mirror.

Unless the proportions do not allow it, true-to-scale reducing or enlarging can also be carried out with the aid of the tracing instrument called the pantograph. When copying, the crayon or pencil inserted in the unequally long feet of the device reproduces the desired contours on the selected scale.

Most of these aids were thus used in normal studio practice and for the preparation of certain applied drawings. Equally practical, but useful only for closely circumscribed tasks, were elliptic compasses, curved rulers, and stencils, particularly for ornamental and decorative purposes. Only a few present-day artists use stencils or simple blocks with a given shape in larger scale composition, in order to obtain the effect of repetition, often in an arbitrary use, in "alienating" technique and colour.

Mechanically produced drawings

Mechanically produced drawings such as typewriter sketches, computer drawings, and oscillograms, all of which can bring forth unusual and attractive results, nevertheless do not belong to the topic because they lack the immediate creativity of the art drawing.

APPLIED DRAWINGS

Applied and technical drawings differ in principle from art drawings in that they record unequivocally an objective set of facts and on the whole disregard aesthetic considerations. The contrast to the art drawing is sharpest in the case of technical project drawings, the purpose of which is to convey not so much visual plausibility as to give exact information that makes possible the realization of an idea. Such plans for buildings, machines, and technical systems are not instantly readable because of the orthogonal (independent) projection, the division into separate planes of projection, and the use of symbols. Prepared as a rule with such technical aids as ruler and compass, they represent a specialized language of their own, which must be learned. For topographic (detailed delineation of the features of a place) and cartographic (map-making) drawings, too, a special terminology has developed that above all systematizes spatial representations, making them intelligible to the expert with the aid of emblems and symbols.

Equally far removed from any claim to artistic standing are most illustrations serving scientific purposes, the aim of which is to record as objectively as possible the characteristic and typical features of a given phenomenon. The systematic drawings, used especially in the natural sciences to explain a system or a function, resemble plans; descriptive and naturalistic illustrations, on the other hand, approach the illusionistic plausibility of visual experience and can attain an essentially artistic character. A good many artists have drawn scientific illustrations, and their works—the botanical and zoological drawings of the Swiss Merian family in the 17th and 18th centuries, for example—are today more esteemed for their artistic than for their documentary value.

Of a similarly ambivalent nature is the illustrative drawing that perhaps does not go beyond a simple pictorial rendition of a literary description but because of its specific formal execution may still satisfy the highest artistic demands. Great artists have again and again illustrated Bibles, prayerbooks, novels, and literature of all kinds. Some of the famous examples are Botticelli's illustrations for Dante's *Divine Comedy* and Dürer's marginal illustrations for the emperor Maximilian's prayer book. Some artists have distinguished themselves more as illustrators than as autonomous draftsmen, as for example the 18th-century German engraver Daniel Chodowiecki, the 19th-century caricaturist Honoré Daumier, the 19th-century satiric artist Wilhelm Busch, and the 20th-century Austrian illustrator Alfred Kubin.

Clearly connected with illustrative drawing is caricature, which, by formally overemphasizing the characteristic traits of a person or situation, creates a suggestive picture that—precisely because of its distortion—engraves itself on the viewer's mind. This special kind of drawing was done by such great artists as Leonardo, Dürer, and the 17th-century artist Gian Lorenzo Bernini and by draftsmen who, often for purposes of social criticism, have devoted themselves wholly to caricaturing, such as the 18th-century Italian Pier Leone Ghezzi, the 19th-century Frenchman Grandville (professional name of Jean-Ignace-Isidore Gérard), and Daumier.

Copyright reserved



"Five Grotesque Heads," pen and ink drawing by Leonardo (1452–1519). In the Royal Library, Windsor Castle. 26 × 20.5 cm.

Modern cartoons

From such overdrawn types developed continuous picture stories that could dispense to a considerable extent with the explanatory text. Modern cartoons are based on these picture stories. Through the formally identical treatment of peculiar types, these drawings acquire an element of consecutiveness that, by telling a continuing story, adds a temporal dimension to two-dimensional drawing. This element is strongest in trick drawings that fix on paper, in brief segments of movement, invented creatures and phenomena that lack all logical plausibility; a rapid sequence of images (leafing through the pages, seeing it projected on the screen) turns the whole into apparent motion, the fundamental process of animation. The artistic achievement, if any, lies in the original invention; its actual realization is predetermined and sometimes carried out by a large and specialized staff of collaborators, often with the aid of stencils and traced designs. Moreover, since the final result is partially determined by the mechanical multiplication, an essential criterion of drawing—the unity of work and result—does not apply.

SUBJECT MATTER OF DRAWING

Anything in the visible or imagined universe may be the theme of a drawing. In practice, however, by far the greatest number of art drawings in the Western world deal with the human figure. This situation springs from the close bond between drawing and painting: in sketches, studies, and compositions, drawing prepared the way for painting by providing preliminary clarification and some formal predetermination of the artist's concept of a given work. Many drawings now highly regarded as independent works were originally "bound," or "latent," in that they served the ends of painting or sculpture. Yet, so rounded, self-contained, and aesthetically satisfying are these drawings that their erstwhile role as handmaidens to the other pictorial arts can be reconstructed only from knowledge of the completed work, not from the drawing itself. This situation is especially true of a pictorial theme that acquired, at a relatively early stage, an autonomous rank in drawing itself: the portrait.

Portraits. Drawn 15th-century portraits—by Pisanello or Jan van Eyck, for example—may be considered completed pictorial works in their concentration, execution, and distribution of space. The clear, delicately delineated representation follows every detail of the surface, striving for realism. The profile, rich in detail, is preferred; resembling relief, it is akin to the medallion. Next in prominence to the pure profile, the three-quarter profile, with its more spatial effect, came to the fore, to remain for centuries the classic portrait stance.

The close relationship to painting applies to practically all portrait drawings of the 15th century. Even so forceful a work as Dürer's drawing of the emperor Maximilian originated as a portrait study for a painting. At the same time, however, some of Dürer's portrait drawings clearly embody the final stage of an artistic enterprise, an ambivalence that can also be observed in other 16th-century portraitists. The works of Jean and François Clouet in France and of the younger Hans Holbein in Switzerland and even more markedly in England in the same century bestowed an autonomy on portrait drawing, especially when a drawing was completed in chalk of various colours. The choice of the softer medium, the contouring, which for all its exactitude is less severely self-contained, and the more delicate interior drawing with plane elements gives these drawings a livelier, more personal character and accentuates once more their proximity to painting.

Chalk and pastel portraits

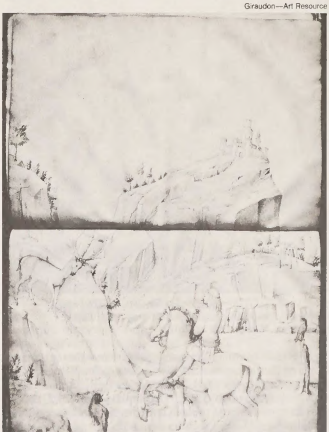
In polychromatic chalk technique and pastel, portrait drawing maintained its independence into the 19th century. In the 18th century, Quentin de La Tour, François Boucher, and Jean-Baptiste Chardin—all of these artists from France—were among its chief practitioners, and even Ingres, living in the 19th century, still used its technique. In pastel painting, the portrait outweighed all other subjects.

In the choice of pose, type, and execution, portrait painting, like other art forms, is influenced by the general stylistic features of an epoch. Thus, the extreme pictorial attitude of the late Baroque and Rococo was followed by a severer conception during Neoclassicism, which preferred monochrome techniques and cultivated as well the special form of the silhouette, a profile contour drawing with the area filled in in black. Unmistakably indebted to their 15th-century predecessors, the creators of portrait drawings of the early 19th century aimed once more at the exact rendition of detail and plastic effects gained through the most carefully chosen graphic mediums: the thin, hard pencil was their favourite instrument, and the silverpoint, too, was rediscovered by the Romantics.



"Portrait of Marguerite de Valois," chalk drawing by François Clouet, *c.* 1559. In the Musée Condé, Chantilly, France. 30.1 × 21.1 cm.
Giraudon—Art Resource

More interested in the psychological aspects of portraiture, late 19th- and 20th-century draftsmen prefer the softer crayons that readily follow every artistic impulse.



"St. Hubert," two pages from a sketchbook by Jacopo Bellini (*c.* 1400–*c.* 1470), pen and ink over chalk or lead point. In the British Museum. 67.2 × 41.5 cm.
Giraudon—Art Resource

The seizing of characteristic elements and an adequate plane rendition weigh more heavily with them than realistic detail. Mood elements, intellectual tension, and personal engagement are typical features of the modern portrait and thus also of modern portrait drawing, an art that continues to document the artist's personal craftsmanship beyond the characteristics of various techniques.

Landscapes. As early as the 15th century, landscape drawings, too, attained enough autonomy so that it is hard to distinguish between the finished study for the background of a particular painting and an independent, self-contained sketched landscape. Already in Jacopo Bellini's 15th-century sketchbooks (preserved in albums in the British Museum and the Louvre), there is an intimate connection between nature study and pictorial structure; in Titian's studio in the 16th century, landscape sketches must have been displayed as suggestions for pictorial backgrounds.

But it was Dürer who developed landscape as a recollected image and autonomous work of art, in short, as a theme of its own without reference to other works. His watercolours above all but also the drawings of his two Italian journeys, of the surroundings of Nürnberg, and of the journey to the Netherlands, represent the earliest pure landscape drawings. Centuries had to pass before such drawings occurred again in this absolute formulation.

Landscape elements were also very significant in 16th-century German and Dutch drawings and illustrations. The figurative representation, still extant in most cases, is formally quite integrated into the romantic forest-and-meadow landscape, particularly in the works of the Danube School—Albrecht Altdorfer and Wolf Huber, for example. More frequently than in other schools, one finds here carefully executed nature views. In the Netherlands, Pieter Bruegel drew topographical views as well as free landscape compositions, in both cases as autonomous works.

By courtesy of the Albertina, Vienna



"The Sacrifice of Isaac," by Albrecht Altdorfer (1480?–1538), pen and ink, heightened with white, on gray-green paper. In the Albertina, Vienna. 20.8 × 17.5 cm.

17th-century landscape drawing

In the 17th century, the nature study and the landscape drawing that grew out of it reached a new high. The landscape drawings of the Accademia degli Incamminati (those of Domenichino, for example) combined classical and mythological themes with heroic landscapes. The Frenchman Claude Lorrain, living in Rome, frequently worked under the open sky, creating landscape drawings with a hitherto unattained atmospheric quality. This type of cultivated and idealized landscape, depicted also by Poussin and other Northerners residing in Rome (they were called Dutch Romanists in view of the fact that so many artists from the Netherlands lived in Rome, their drawings of Italy achieving an almost ethereal quality), is in contrast with the unheroic, close-to-nature concept of landscape held primarily by the Netherlanders when depicting the landscape of their native country. All landscape painters—their landscape paintings a specialty that was strongly represented in the artistically specialized Low Countries—also created independent landscape drawings (Jan van Goyen and Jacob van Ruisdael and his uncle and cousin, for example), with Rembrandt again occupying a special position: capturing the characteristics of a region often with only a few strokes, he enhanced them in such manner that they acquire monumental expressive power even in the smallest format. In 18th-century Italy, the topographically faithful landscape drawing gained in importance with the advent of the *Vedutisti*, the purveyors of "views," forming a group by themselves (among them, Giambattista Piranesi and Canaletto [Giovanni Antonio Canal]) and often working with such optical aids as the graticulate frame and camera obscura. Landscape drawings of greater artistic freedom, as well as imaginary landscapes, were done most successfully by some French artists, among them Hubert Robert; pictorially and atmospherically, these themes reached a second flowering in the brush-drawn landscapes of such English artists as Turner and Alexander Cozens, whose influence extends well into the 20th century.

Given their strong interest in delineation, the 18th-century draftsmen of Neoclassicism and, even more, of Romanticism observed nature with topographical accuracy. As a new "discovery," the romantically and heroically exaggerated Alpine world now took its place in the artist's mind alongside the arcadian view of the Italian landscape.

Landscape drawings and even more, watercolours, formed an inexhaustible theme in the 19th century. The French artist Jean-Baptiste-Camille Corot and, toward the end of the century, Cézanne and van Gogh, were among the chief creators of landscape drawings. While landscapes form part of the work of many 20th-century draftsmen, the genre as such takes second place to general problems of form, in which the subject is merely treated as starting point.

Figure compositions and still lifes. Compared to the main themes of autonomous drawing—portraiture and landscape—all others are of lesser importance. Figure compositions depend greatly on the painting of their time and are often directly connected with it. There were, to be sure, artists who dealt in their drawings with the themes of monumental painting, such as the 17th-century engraver and etcher Raymond de La Fage; in general, however, the artistic goal of figure composition is the picture, with the drawing representing but a useful aid and a way station. Genre scenes, especially popular in the 17th-century Low Countries (as done by Adriaen Brouwer, Adriaen van Ostade, and Jan Steen, for example) and in 18th-century France and England, did attain some independent standing. In the 19th century, too, there were drawings that told stories of everyday life; often illustrative in character, they may be called "small pictures," not only on account of the frequently multicoloured format but also in their artistic execution.

Still lifes can also lay claim to being autonomous drawings, especially the representations of flowers, such as those of the Dutch artist Jan van Huysum, which have been popular ever since the 17th century. Here, again, it is true that a well-designed arrangement transforms an immediate nature study into a pictorial composition. In some of these compositions the similarity to painting is very strong; the pastels of the 19th- and 20th-century artist Odilon Redon, for instance, or the work of the 20th-century German Expressionist Emil Nolde, with its chromatic intensity, transcend altogether the dividing line between drawing and painting. In still lifes, as in landscapes, autonomous principles of form are more important to modern artists than the factual statement.

Fanciful and nonrepresentational drawings. Drawings with imaginary and fanciful themes are more independent of external reality. Dream apparitions, metamorphoses, and the entwining of separate levels and regions of real-





Contrasting approaches to the representation of landscape.
(Top) "Pastoral Landscape," by Claude Lorrain, pen with brown and gray brown wash, from the *Liber Veritatis* (78), 1644. In the British Museum. 18.5 × 26 cm. (Bottom) "House amid Trees on the Bank of a River," by Rembrandt (1606–69), pen and black ink, India ink wash, on brown coloured paper. In the British Museum. 40.6 × 59.2 cm.
By courtesy of the Trustees of the British Museum: photographs: J.R. Freeman & Co. Ltd.

ity have been traditional themes. The late 15th-century phantasmagoric works of Hieronymus Bosch are an early example. There are allegorical peasant scenes by the 16th-century Flemish artist Pieter Bruegel and the carnival etchings of the 17th-century French artist Jacques Callot. Others whose works illustrate what can be done with drawing outside landscape and portraiture are: the 18th-century Italian engraver Giambattista Piranesi, the 18th-century Anglo-Swiss artist Henry Fuseli, the 19th-century English illustrator Walter Crane, the 19th-century French Symbolist artist Gustave Moreau, and the 20th-century Surrealists.

Nonrepresentational art

Nonrepresentational art, with its reduction of the basic elements of drawing—point, line, plane—to pure form, offered new challenges. Through renunciation of associative corporeal and spatial relationships, the unfolding of the dimensions of drawing and the structure of the various mediums acquire new significance. The graphic qualities of the line in the plane as well as the unmarked area had already been emphasized in earlier times—for example, in the *grotteschi* of Raphael in the 16th century (the fanciful or fantastic representations of human and animal forms often combined with each other and interwoven with representations of foliage, flowers, fruit, or the like) and in calligraphic exercises such as moresques (strongly stylized linear ornament, based on leaves and blossoms)—but mostly as printing or engraving models for the most disparate decorative tasks (interior decoration, furniture, utensils, jewelry, weapons, and the like).

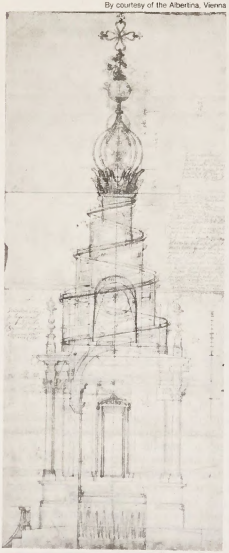
Artistic architectural drawings. There is one field in which drawing fulfills a distinct function: artistic architectural drawings are a final product as drawings, differing from the impersonal, exact plans and designs by the same "handwriting" character that typifies art drawings. In many cases, no execution of these plans was envisaged; since the early Renaissance, such ideal plans have been drawn to symbolize, in execution and accessories, an abstract content. Despite the often considerable exactitude with which the plans are drawn, the personal statement



"Design for a Wall Panel," by Giambattista Piranesi (1720–78), pen with brown ink, brown wash, over black chalk. In the Pierpont Morgan Library, New York. 28.8 × 28.2 cm.
By courtesy of the Pierpont Morgan Library, New York

predominates in the flow of the line. This personal note clearly identifies the drawings of such artists and architects as Albrecht Altdorfer, Leonardo, Michelangelo, Bernini, Francesco Borromini, and Piranesi. Also distinct from the ground-plan type of architectural drawing are the art drawings of autonomous character created by such 20th-century architects as Erich Mendelsohn and Le Corbusier.

By courtesy of the Albertina, Vienna



"S. Ivo della Sapienza," architectural drawing by Francesco Borromini, c. 1642–60. In the Albertina, Vienna. 41 × 26.7 cm.

History of drawing

WESTERN

As an artistic endeavour, drawing is almost as old as mankind. In an instrumental, subordinate role, it developed along with the other arts in antiquity and the Middle Ages. Whether preliminary sketches for mosaics and murals or architectural drawings and designs for statues and reliefs within the variegated artistic production of the Gothic medieval building and artistic workshop, drawing as a nonautonomous auxiliary skill was subordinate to the other arts. Only in a very limited sense can one speak of centres of drawing in the early and High Middle Ages; that is, the scriptoria of the monasteries of Corbie and Reims in France, as well as those of Canterbury and Winchester in England, and also a few places in southern Germany, where various strongly delineatory (graphically illustrated) styles of book illumination were cultivated.

14th, 15th, and 16th centuries. In the West, the history of drawing as an independent artistic document began toward the end of the 14th century. If its development was independent, however, it was not insular. Just as the greatest draftsmen have been for the most part also distinguished painters, illustrators, sculptors, or architects, so the centres and the high points of drawing have generally coincided with the leading localities and the major epochs of the other arts. Moreover, the same stylistic phenomena have been expressed in drawing as in other art forms. Indeed, drawing shares with other art forms the characteristics of individual style, period style, and regional features. Drawing differs, however, in that it interprets and renders these characteristics in terms of its own unique mediums.

Drawing as an independent art form

Drawing became an independent art form in northern Italy, at first quite within the framework of ordinary studio activity. But with nature studies, copies of antiques, and drafts in the various sketchbooks (those of Giovannino de'Grassi, Antonio Pisanello, and Jacopo Bellini, for example), the tradition of the Bauhütten studio workshop changed to individual work: the place of "exempla," models, reproduced in formalized fashion was now being taken by subjectively probing and partially creative drawings. In the early 15th century the international Soft Style of the period still largely predominated over the draftsman's individual "handwriting." At mid-century, however, the differentiation of drawing style according to region and the artist's personality set in. Essential criteria, destined to remain characteristic for generations, begin to strike the eye.

In drawing produced north of the Alps, the characteristic features lie in the tendency to pictorial compactness and precise execution of detail. Many painters produced individual drawings, but the most notable draftsmen are the otherwise unidentified 15th-century German Master of the Housebook and his contemporary Martin Schongauer. Both of these artists were also major copperplate engravers, so that it is not always easy to determine whether the work is a preliminary sketch or an independent drawing.

In Italian Renaissance drawings, of which there are a great many, the diverging stylistic features of the various artistic regions were particularly evident. What they had in common was the overwhelming importance of the sketch and the study, in contrast to the far rarer finished drawings. The formal and thematic connection with painting is very close even when it was not a question of preliminary drawings. The draftsmen of Venice and northern Italy preferred an open form with loose and interrupted delineation in order to achieve even in drawing the pictorial effect that corresponded to their painters' imagination.

In central Italy, on the other hand, and especially in Florence, it was the clear contour that predominated, the closed and firmly circumscribed form, the static and plastic character. Corresponding to the functional purpose of drawing, the individual artists' studios (which, as was the case with the Medicis' Academy of St. Mark, also had to engage in general educational and humanistic investigations) formed the most significant centres of art drawing. In these large studios, drawing served not only for the probing realization of creative ideas, it was not only study and mediator between the conception and the master's finished work; it functioned also as teaching aid for the

assistants who worked with the master and as a vehicle for the formation and preservation of an individual workshop tradition. Although Leonardo's scientific interests were expressed in a large number of drawings, his ideal concept of the human figure is much more frequently preserved in the drawings of his collaborators and successors than in his own. Raphael and Michelangelo were also outstanding draftsmen. Each of them used drawing in order to allow his thoughts about individual works to mature; each had a highly personal drawing style, the one with a soft and rounded stroke, the other with a sculptor's intermittent and powerful stroke. Probably a great deal of drawing was done in Raphael's studio, especially if only for the preparation of the engravings after Raphael's compositions. From Michelangelo's hand came the first so-called connoisseur drawings that are esteemed as a personal document. They are the precursors of the collector's drawings that began in the later 16th century (autonomous works, destined for collections).

Northern European drawing

North of the Alps the autonomy of drawing was championed in the first instance by Dürer, an indefatigable draftsman who mastered all techniques and exercised an enduring and widespread influence. The delineatory constituent clearly predominates even in his paintings. This corresponds to the general stylistic character of 16th-century German art, within which Matthias Grünewald, with his freer, broader, and therefore more pictorial style of drawing, and the painters of the Danube school, with their ornamentalizing and agitated stroke, represent significant exceptions. In their metamorphosing of the perceived reality into drawings, the landscapes of Altdorfer and Wolf Huber in particular are astonishing documents of a feeling for nature that might almost be called Romantic.

Soberer, incredibly compact in their pictorial concept and yet akin to the Renaissance in their objective viewing, were the portrait drawings of Hans Holbein, the Younger, whose sojourns in 16th-century England proved stimulating to other artists as well. Similar, if less personal than Holbein because of the stricter linearity of their work, were the drawings of the French portraitists Jean and François Clouet. In the Low Countries, where they were combined with the idealized image of Italy (as in the drawings of Lucas van Leyden), Dürer's methods gained lasting popularity in the landscape drawings and studies "after life" by Pieter Bruegel the Elder.

Drawing acquired a pivotal significance in the period of Mannerism (*c.* 1525–1600), both as a document of artistic invention and as a means of its realization. Jacopo Pontormo in Florence, Parmigianino in northern Italy, and Tintoretto in Venice used point and pen as essential and spontaneous vehicles of expression. Their drawings were clearly related to their painting, both in content and in the graphic method of sensitive contouring and daringly drawn foreshortening.

17th, 18th, and 19th centuries. In the early 17th century, Jacques Callot rose to prominence in French art: gifted as a draftsman above all, he recorded with the pen his clever inventions and great picture stories, primarily in bold abbreviations.

The importance of drawing for an artist's growth and the widening of his horizon is attested also by the work of Peter Paul Rubens, whose studies and sketches make up an integral part of his creative achievement. In order to disseminate his pictorial themes and concept of form, he maintained his own school for draftsmen and engravers. Among the circle of Flemings around him, Jacob Jordaens and Sir Anthony Van Dyck are notable as draftsmen with a style of their own.

Hercules Seghers was among the most fascinating artists of the 17th century, a creator of drawn and etched landscapes that he continued to rework while experimenting with printing processes. From the point of view of technique and form, he was important for the greatest artist of Holland, Rembrandt. Seghers combined great inventiveness, especially in his interpretations of Old Testament motifs, and broad mastery of all the techniques of drawing. In his studio, too, drawing was emphasized as a teaching aid and a means of formal experimentation.

Most Dutch painters of the 17th century, such as the van de Velde family, Brouwer, van Ostade, Pieter Saenredam, and Paulus Potter, were also industrious draftsmen who recorded their special thematic concerns in drawings that were largely completed. Beyond serving as preparation for paintings, these were regarded as autonomous works representing the final stage of the creative process.

In 17th-century Italy, drawing by way of artistic practice and experimentation became established in the academies, especially in Bologna. More significant, however, was the continuing development of landscape drawing, as initiated by the brothers Agostino and Annibale Carracci and articulated further by Domenichino and Salvator Rosa. The French artist Claude Lorrain so developed the landscape drawing of the Roman countryside that it became almost a genre of its own; in his works, which were often intended for sale, nature study and an idealized pictorial concept are uniquely merged. In detailed studies directly before the object, he achieved a timeless validity. Like Lorrain, Poussin also drew under the open sky. Using various techniques, he combined realistic experiences and humanistic concepts in idealizing compositions; the figures and scenes of which are harmoniously integrated into a spacious landscape. This open-air painting and drawing was practiced also by some other artists who spent a considerable time in Rome—the Dutch artists Jan Asselijn, Claes Berchem, Karel Dujardin, and Adam Pijnacker, for example. For most southern European artists of the 17th century, however, drawing was a mere stage in the creation of a painting.

17th-century Italian landscape drawing

Antoine Watteau, too, did drawings to "keep his hand in" for his painting, although he did so with an independence that led him far beyond the immediate occasion. Most figures in the paintings from various periods of his career were based on earlier drawings. In the grand scale of his form and the attention paid to pictorial elements, he carried on in the manner of Rubens, combining it with the light esprit of the 18th century. The leading position of French art in the first half of that century was confirmed by the achievements of Boucher, Fragonard, Hubert Robert, and Gabriel de Saint-Aubin, whose drawings include figure studies, genre-like works, and landscapes.

In contrast to the French draftsmen who brought about a flowering of the *à trois crayons* method on tinted paper, some artists created similar landscapes with pen and brush but with greater objective abbreviation. Mention must here be made of Venice, with the Giovanni Battista Tiepolo family, whose expansively conceived pen drawings, washed with a broad brush, call forth the kind of luminaristic effect that Francesco Guardi also used for landscape studies and imaginary scenes. These had been preceded by Canaletto's views of Venice, composed more severely as far as tectonic (constructional) detail is concerned but nonetheless the first examples of this form of the landscape capriccio, or fantasy. The architect Giambattista Piranesi made his name primarily as a draftsman who recorded views of Rome; above all, in his drawings of architecture and eerie vaults ("Carceri"), he left behind a body of work of great intellectual and formal forcefulness.

The Spanish painter Goya, at the very end of the 18th and in the beginning of the 19th century, was in advance of his time in the way in which he handled his themes. Forming an odd contrast to the court-painter's pictures, his brush-and-sanguine drawings are rather more closely tied to his cycles of etchings. He combined the luministic effects of Tiepolo's drawings with the dramatic impact of a Rembrandt chiaroscuro.

Also at the turn of the 19th century is an artist whose main work was that of a draftsman: the English caricaturist and social satirist Thomas Rowlandson, who produced colourful and distinctive watercolours. The late 18th and, even more, the early 19th century produced a drawing style that, in accordance with both the Neoclassical and the Romantic ideal, emphasized once more the linear element. In Ingres, idealistic Neoclassicism found an exemplary expression of strict linearity, and the pencil drawing became a downright classical form. The Nazarenes and Romantics in Rome and the Alpine region (Joseph Anton Koch, the brothers Friedrich and Ferdinand Olivier, and Julius Schnorr von Carolsfeld) as well as those in

north Germany (Philipp Otto Runge and Caspar David Friedrich) were more lyrical but equally rigorous in the use of the hardpoint; after a long time, they were the first northern artists to have made a significant contribution to the history of drawing. Among 19th-century artists, the emphasis on delineation was characteristic also of Moritz von Schwind in Germany and John Millais in England. (In the Neoclassical phase of the 20th century it was renewed, in a more open and "handwriting" fashion, by Thomas Eakins in the United States as well as by Picasso, Matisse, and Amedeo Modigliani in France.) The drawings of Delacroix, while preserving plastic qualities, show a broader stroke and are thus more pictorial. Daumier, active in all mediums primarily as a draftsman, utilized pictorial chiaroscuro effects in forcible statements of social criticism.

19th-century France

France continued to be a leading centre of the art of drawing, a form that was given a very personal note in each case in the works of Degas, Toulouse-Lautrec, van Gogh, and Cézanne. The line—the common point of departure for all of the above-mentioned artists—did not disappear until Seurat's plane shading, done in the Pointillist manner.

Modern. Except for a few stylistic currents such as Tachism (paintings consisting of irregular blobs of colour), drawing is represented in the work of practically all 20th-century artists; it is as international as modern art itself. As the other arts have become nonrepresentational, thus attaining autonomy and formal independence in relation to external reality, drawing is more than ever considered an autonomous work of art, independent of the other arts; and the sketch, study, and project—that is, the drawing as a stage in the genesis of works of sculpture, painting, and architecture—have greatly diminished in importance. Some schools and individual artists as well have concentrated on drawing and in very individualistic ways. The German Expressionists, for instance, developed especially emphatic forms of drawing with powerful delineation and forcible and hyperbolic formal description; notable examples are the works of Ernst Barlach, Käthe Kollwitz, Alfred Kubin, Ernest Ludwig Kirchner, Karl Schmidt-Rottluff, Max Beckmann, and George Grosz. In the artists' group *Der Blaue Reiter* (The Blue Rider), Wassily Kandinsky was foremost in laying the groundwork for a new evaluation of the nonrepresentational line. Paul Klee's lyrically sensitive drawings, carried out in a pen technique of unheard-of sublimity, represent a high point of modern drawing. In France, drawing plays a major role, especially in the work of the painters of the École de Paris (School of Paris), such as Pierre Soulages and Hans Hartung, who consider the line, the framework of lines, and the network of lines, as primary manifestations of form. Wols (Alfred Otto Wolfgang Schulze) and also the English artist Graham Sutherland may actually be called spiritual draftsmen who put their faith in the magic of the line. Finally, drawing occupies a considerable place in the work (including all its variants of style and form) of Picasso, once again a man who knew how to make use of its manifold technical possibilities. One is surely justified in calling him the greatest draftsman of the 20th century and one of the greatest in the history of drawing. (H.R.H.)

EASTERN

Drawings of China

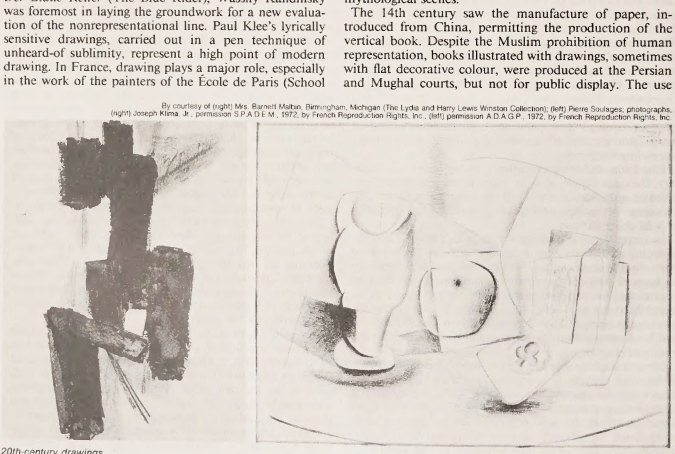
Some form of monochromatic brush drawing with ink may have been practiced in China as early as the 2nd millennium BC; but the earliest pictorial work is in lacquer or on bronze vessels, contemporaneous with Alexander the Great (ruled 336–323 BC). It relies on contour and silhouette, with men and animals depicted in horizontal registers (levels, one above the other) reminiscent of Egyptian and Mediterranean work. The extent of any mutual influence between East and West cannot yet be determined. Under the Tung (Eastern) Han dynasty (AD 25–220) wall paintings, linear in character, were produced in fresco (wet plaster) and secco (dry). Only in the Wei (386–534/35) and T'ang (618–907) dynasties did the true character of Chinese drawing on silk or paper emerge. In the 7th century, the characteristic albums (*ts'e-yeh*) of drawings appear.

No distinction was made between drawing and painting because all Chinese pictorial art was fundamentally graphic. The artist worked with the fine point of the brush on paper or silk laid horizontally on a table. Work in pure outline was called *pai-miao*; ink applied in splashes, *p'o-mo*. Colour was used sparingly or not at all. The final work was not made direct from nature.

Hindu and Buddhist paintings at Ajantā in India and also in Ceylon reveal the essential quality in all Indian art: emphasis on a flowing, rhythmic contour to express movement and gesture. Drawings on palm leaf of the 11th century are similarly based on the use of line to depict mythological scenes.

The 14th century saw the manufacture of paper, introduced from China, permitting the production of the vertical book. Despite the Muslim prohibition of human representation, books illustrated with drawings, sometimes with flat decorative colour, were produced at the Persian and Mughal courts, but not for public display. The use

By courtesy of (right) Mrs. Barnett Malbin, Birmingham, Michigan (The Lydia and Harry Lewis Winston Collection); (left) Pierre Soulages; photographs, (right) Joseph Klima, Jr., permission S.P.A.D.E.M., 1972, by French Reproduction Rights, Inc., (left) permission A.D.A.G.P., 1972, by French Reproduction Rights, Inc.



20th-century drawings.
(Left) Drawing by Pierre Soulages, walnut stain and graphite, 1950. In the collection of the artist. 65 × 50 cm. (Right) "Still Life with Glass, Apple, Playing Card, and Package of Tobacco," pencil drawing by Pablo Picasso, 1913. In the Lydia and Harry Lewis Winston Collection, Birmingham, Michigan. 23.8 × 31.0 cm.





Eastern drawing.
(Left) "The Merchant and the Ascetic," Mughal drawing with colours, early 17th century. In the Smithsonian Institution, Freer Gallery of Art, Washington, D.C. 10.3 × 9.3 cm. (Right) "Woodcutter Gazing at Waterfall" (detail), by Hokusai, ink and colour on paper scroll, 1798. In the Stanford University Museum of Art, California. 29.7 × 39.0 cm.
By courtesy of (left) the Smithsonian Institution, Freer Gallery of Art, Washington, D.C., (right) the Stanford University Museum of Art, California, Ikeda Collection

of a precise and expressive line constituted the basis for Persian and Indian (both Mughal and Rājput) miniature paintings, which show people in landscape or in relation to buildings.

Japanese art tended to follow that of China until the early 19th century, when the popular colour print was introduced. In the graceful feminine gestures of Utamaro's work, the Oriental love of flowing contour is manifest, his lines varying in width and density. Hokusai's drawings of social life in a humorous, almost grotesque vein reveal his complete command of the expressive line. (Ed.)

BIBLIOGRAPHY. JOSEPH MEDER, *The Mastery of Drawing*, trans. and rev. by WINSLOW AMES, 2 vol. (1978; originally published in German, 1919; 2nd ed., 1923), a voluminous work that remains the basic treatment of the history and techniques of drawing. Another treatment, more concise in every respect, is HEINRICH LEPORINI, *Die Künstlerzeichnung*, 2nd ed. (1955). ARTHUR E. POPHAM published an introduction to drawing in *A Handbook to the Drawings and Watercolours in the Department of Prints and Drawings of the British Museum* (1939), based on the ample materials held by the British Museum. WALTER KOSCHATZKY, *Die Kunst der Zeichung: Technik, Geschichte, Meisterwerke* (1977), is a survey of the history, functions, and techniques of drawing, from the beginnings to modern art, based on excellent examples chosen mainly from the Graphische Sammlung Albertina in Vienna. CHARLES DE TOLNAY in *History and Technique of Old Master Drawings* (1943, reprinted 1972); and JAMES WATROUS in *The Craft of Old-Master Drawings* (1957), deal, from different points of view, with the history and techniques of the old masters; while HERIBERT HUTTER in *Drawing: History and Technique* (1968; originally published in German, 1966), stresses the artistic function of drawing and includes modern works. DANIEL M. MENDELOWITZ, *Mendelowitz's Guide to Drawing*, 3rd ed. rev. by DUANE A. WAKEHAM (1982), provides a historical résumé, with reference to the artistic elements and technical means of drawing; in the supplement to the 1st ed., *Drawing: A Study Guide* (1967), he offers practical instructions for drawing techniques and their application; as does ROBERT BEVERLY HALE in *Drawing Lessons from the Great Masters* (1964, reprinted 1974). JAKOB ROSENBERG illustrates the possibilities of drawing in *Great Draughtsmen from Pisanello to Picasso*, rev. ed. (1974), with samples from the works of eight great artists. *Great Drawings of All Time*, ed. by IRA MOSKOWITZ and VICTORIA THORSTON, 5 vol. in 6 (1962–79), contains a summary with comments by leading authorities. M.W. EVANS, *Medieval Drawings* (1969), is useful for the early history of the art of drawing; PAUL J. SACHS, *Modern Prints and Drawings: A Guide to a Better Understanding of Modern Draughtsmanship* (1954), for more recent developments. HERMANN BOEKHOFF and FRITZ WINZER, *Das grosse Buch der Graphik* (1968), gives the history of the 24 best known collections, with comments by the various curators and the basic catalog of each collection. Interesting information can be found in catalogs of many exhibitions and collections, such as BERNICE ROSE, *Drawing Now* (1976), which discusses contemporary types of drawing. The number of detailed investigations in regard to individual countries, periods, and artists is too large to be listed in this bibliography. One that can be especially recommended, however, is EDWARD J. OLSZEWSKI, *The Draftsman's Eye: Late Italian Renaissance Schools and Styles* (1981). LUIGI GRASSI, *Storia del disegno* (1947), is very valuable for the role of drawing in the historical theories of art, including the elucidation of the original sources for further study. Unsurpassed in method and fundamental for an intensive study of this subtle theme is BERNHARD DEGENHART'S essay "Zur Graphologie der Handzeichnung," in *Jahrbuch der Hertziana*, vol. 1 (1937).

(H.R.H.)

Dress and Adornment

It is not possible, in an article of this length, to discuss the immense variety of garments and ornamentation worn by human beings throughout the world from prehistoric until modern times. Therefore, the following types of attire will not be treated here: ecclesiastical dress; military dress; academic, trade, or professional dress; and the national or regional costumes of peasant or primitive peoples. What will be considered in some detail is the chronological development of fashionable dress and decoration—that is, the attire selected and adopted by the leading members of a society. This is not an arbitrary decision but rather one reflecting the long-standing fact that in any group of people—whether constituting a small community or a great, influential nation—it is those members with wealth and, consequently, power at any given time who influence, and even dictate, the fashions to other, lesser members. The discussion does not concentrate solely on clothing but also covers, when appropriate, features of coiffure, head coverings, footwear, accessories, and cosmetic beautification. In addition, the nature and purposes of dress and some of the specific social, political, economic, geographic, and technological factors influencing changes in fashion, as well as the uses, forms, styles, materials, and techniques that have characterized jewelry throughout its 40,000-year history, are also treated.

For coverage of related topics in the *Macropædia* and *Micropædia*, see the *Propædia*, section 629, and the *Index*.

This article is divided into the following sections:

Dress 478
- The history of Middle Eastern and Western dress 478
  - Ancient Egypt
  - Mesopotamia
  - The Aegean: Minoan and Mycenaean dress
  - Ancient Greece
  - Etruria
  - Ancient Rome
  - Ancient nonclassical Europe
  - The pre-Columbian Americas
  - The Middle East from the 6th century
  - The Byzantine Empire
  - Medieval Europe
  - Europe, 1500–1800
  - Colonial America
  - The Ottoman Empire
  - Europe and America: 19th and 20th centuries
- The history of Eastern dress 500
  - China
  - Japan
  - Korea
  - South Asia
- The nature and purposes of dress 504
  - Display of the human physique
  - Government regulation of dress
  - Rebellion
  - Exotica

Jewelry 509
- Materials and methods 509
  - Metals
  - Gems
- The history of jewelry design 512
  - Middle Eastern and Western antiquity
  - Middle Ages
  - Renaissance to modern
  - Non-Western cultures

Bibliography 527

DRESS

The history of Middle Eastern and Western dress

ANCIENT EGYPT

Modern knowledge of ancient Egyptian dress derives from the ample evidence to be seen in the wealth of wall and sarcophagus paintings, in sculpture, and in ceramics; few actual garments have survived. Such illustrative material is depicted clearly and colourfully, but care must be taken in interpreting the designs too literally, partly because the art is frequently stylized but also because the artists were bound by tradition and their representation of dress often lagged far behind the actual changes of fashion.

Egyptian fabrics

The chief textile to have been preserved is linen, which has been found in graves dating to Neolithic times. Flax culture dates from very early times, and, in fact, the Egyptians believed that the gods were clothed in linen before they came to earth. Wool was more rarely employed, and sericulture had not yet extended as far west as Egypt. The technique of using mordants in the dyeing processes was slow to come to Egypt, so most garments were white. Colour was provided by jewelry in which semiprecious stones were widely incorporated. Among the most common types, the characteristic deep, decorative collar, worn by both sexes, was introduced early. These brightly hued bands were made of embroidered and beaded materials and set around the neck and shoulders either on bare skin or on top of a white cape or gown.

Skins of various animals were utilized. These were sometimes simply raw hides, which have survived only rarely, but the Egyptians became skilled at curing the skins to become leather by the tawing method—that is, by the use of alum or salt. Tawing yields a white, stiff leather that may be dyed various colours. Later they adopted the tanning method, employing oak galls for the purpose. Leather was used widely in dress for footwear, belts, and straps.

During the 3,000 years of the Egyptian culture, costume changed comparatively little and very slowly. It remained a draped style of dress, the garments consisting of pieces of material held in place around the body by knots tied in the fabric and by waist belts, sashes, and collars. Little sewing was needed, being confined generally to side seams and, in later years, to armholes. This draped type of dress conformed to that of other civilizations in the Mediterranean and Middle Eastern region such as Greece, Rome, and Mesopotamia but differed from the more Oriental styles of Persia, India, and China, where people wore more fitted, sewn garments based upon coats, tunics, and trousers. Ancient Egyptian dress for both sexes was confined to loincloths, a type of vest or shirt, capes, and robes.

Over the years the style of these garments slowly evolved and became more complex; a greater number were worn either in combination with or on top of one another. During the Old Kingdom (its capital at Memphis), which lasted until about 2130 BC, dress was simple. Men wore a short skirt tied at the waist or held there by a belt. As time passed, the skirt became pleated or gathered. Important people wore in addition a decorative coloured pendant hanging in front from the waist belt as well as a shoulder cape or corselet partly covering their bare torso. A sheath-like gown was typical of feminine attire. This encased the body from the ankles to just below the breasts and was held up by decorative shoulder straps. Woolen cloaks were worn for warmth by men and women.



Figure 1: Woman wearing sheathlike gown held up by shoulder straps, typical of Egyptian dress of the Old and Middle Kingdoms. Painted wood statue from the tomb of Meketre, Dayr al-Baḥrī, Egypt, 11th dynasty (2081–1938 BC). In the Egyptian Museum, Cairo.

Borromeo/Art Resource, New York City

Middle Kingdom styles

Under the Middle Kingdom, based on Thebes, which prospered until about 1600 BC, the masculine skirt could be hip- or ankle-length. More material was now used, making the garment fuller, such fullness being concentrated in the centre front; and the pendants became more elaborate and ornamental. A cape might be draped around the shoulders and knotted on the chest. Late in the period a double skirt was introduced; alternatively, a triangular loincloth might be worn under a skirt.

Hirmer Fotoarchiv, München



Figure 2: Egyptian dress of the New Kingdom, 18th dynasty. King Tutankhamen wearing a double skirt, long and full, with the upper one doubled and gathered in front; Queen Ankhesenamen in a draped robe tied at the breast and leaving the right arm free. Detail from the back of the throne of Tutankhamen (reigned 1333–23 BC). In the Egyptian Museum, Cairo.

The most elaborate dress for both sexes was to be seen under the New Kingdom from about 1539 BC until the Egyptians were conquered successively by the Assyrians (671 BC), the Persians (525 BC), Alexander the Great (332 BC), and finally Rome (30 BC). During these later years Egyptian dress was strongly influenced by that of the conquerors. New Kingdom dress was more complex than theretofore. The garments were of similar type but were composed of larger pieces of material; draping became more complicated and ornamentation richer. A robe or gown was now worn by important persons of both sexes. It consisted of a piece of fabric measuring 5 feet by 4 feet (1.5 metres by 1.2 metres) that was draped and held in place by pins and a waist belt, creating wide, elbow-length sleeves. There were many ways of draping the material, but with most methods all the pleats and folds seemed to be gathered around a single point at the waist. The cape, decorative collar, skirt, and pendant girdle also continued to be worn. Foci of bright colour were provided by the deep collar and pendant apron. Embroidered and carved ornamental motifs included especially the lotus flower, the papyrus bundle, birds in flight, and many geometric forms. Sacred emblems such as the scarab beetle and the asp were worn by priests and royalty.

Children were dressed, as in most of the history of costume everywhere, as miniature versions of their parents, although they are often depicted wearing little at all—not surprising considering the climate of Egypt. Servants also were almost naked, as were labourers in the fields, who are depicted clad only in a loincloth.

Wigs

Heavy wigs or a padding of false hair, worn by both men and women, are known from an early period. They served not only as an adornment but also to protect the wearer's head from the burning rays of the Sun, thus in a way acting as hats. Semicircular kerchiefs, tied by the corners at the nape of the neck under the hair, were sometimes worn to protect the wig on a dusty day. Wigs were dressed in many different ways, each characteristic of a given period; generally speaking, the hair became longer and the arrangement of curls and braids—set with beeswax—more complicated as time went on.

The earliest records indicate that the Egyptians grew hair on their chins. They frizzed, dyed, or hennaed this beard and sometimes plaited it with interwoven gold thread. Later, a metal false beard, or postiche, which was a sign of sovereignty, was worn by royalty. This was held in place by a ribbon tied over the head and attached to a gold chin strap, a fashion existing from about 3000 BC to 1580 BC.

Many people went barefoot, especially indoors, but people of rank are depicted outdoors in sandals made from palm leaves, papyrus, or leather.

Cosmetics were extensively applied by both sexes, and considerable knowledge of their use is available because of the Egyptian custom of burying comforts and luxuries with the dead. Examples both of the cosmetics and of the means of making, applying, and keeping them may be seen in museums, especially in Cairo and London. The Egyptians applied rouge to their cheeks, red ointment to their lips, and henna to their nails and feet, and ladies traced the veins on their temples and breasts with blue paint, tipping their nipples with gold. The chief focus of makeup was the eye, where a green eye shadow (made from powdered malachite) and a black or gray eyeliner was applied; the latter substance, called kohl, was manufactured from, among other materials, powdered antimony, carbon, and oxide of copper.

MESOPOTAMIA

Ancient Mesopotamia was situated in the area of land that is defined by the two great rivers the Tigris and the Euphrates and that is contained within modern Iraq. Several important cultures arose there, their empires waxing and waning successively as well as overlapping in time. Among the most prominent were the Sumerian, the Akkadian, or Semitic, the Assyrian, and the Babylonian.

Sumerian dress

The Sumerian civilization was established before 4000 BC and reached a high level of culture between 2700 and 2350 BC. In early times both sexes wore sheepskin skirts with the skin turned inside and the wool combed into



Figure 3: Skirt of tufted woolen fabric known as *kaunakes*, worn by Ebih-il, an official of the Temple of Ishtar in Mari, Syria. Statue of stone, bitumen, shell, and lapis lazuli, *c.* 2400 BC. In the Louvre, Paris.

Photograph, © Reunion des Musees Nationaux

decorative tufts. These wraparound skirts were pinned in place and extended from the waist to the knees or, for more important persons, to the ankles. The upper part of the torso was bare or clothed by another sheepskin cloaking the shoulders. From about 2500 BC a woven woolen fabric replaced the sheepskin, but the tufted effect was retained, either by sewing tufts onto the garment or by weaving loops into the fabric. Named *kaunakes* by the Greeks, this tufted fabric is shown in all the sculptures and mosaics of the period, as, for example, in the art from the excavations at Ur exhibited in the British Museum in London. At this time, also, long cloaks were worn, and materials for garments and head coverings included felted wool and leather. Men were generally clean-shaven. Both sexes seem to have often worn large wigs, as in ancient Egypt. Metalworking was of a high standard, as may be seen in the elaborate golden jewelry, which was encrusted with semiprecious stones and worn by both sexes: brooches, earrings, hair ornaments, and neck chains.

A different style of dress is evident in Mesopotamian sculptures dating after about 2370 BC. Both men and women were clothed in a large piece of material—most commonly of wool, though later also of linen—draped around the body over a skirt. This garment, similar to a shawl, was characteristically edged with tassels or fringe. The draping varied, but, for men at least, the fabric was arranged so that the fullness was at the rear, leaving the right, or sword, arm free. This newer form of dress had originated from farther north and east and was adopted by the Semitic people of Akkad under Sargon (the dynasty founded by Sargon lasted from *c.* 2334 BC to *c.* 2193 BC) and by the revitalized Sumerian culture in the years 2110–2010 BC.

The dress worn in Mesopotamia by the Babylonians (2105 BC–1240 BC) and the Assyrians (1200 BC–540 BC) evolved into a more sophisticated version of Sumerian and Akkadian styles. Ample evidence of this more elaborate draped costume can be seen in the large relief sculptures of the age. There were two basic garments for both sexes: the tunic and the shawl, each cut from one piece of material. The knee- or ankle-length tunic had short sleeves and a round neckline. Over it were draped one or more shawls of differing proportions and sizes but all generally fringed or tasseled. Broad belts held the shawls in position. Wool was the most frequently used material, in bright or strong colours. Decoration was rich, in allover patterns or in borders, carried out in embroidery or by printing. Motifs were chiefly geometric. Women wore a short skirt as underwear, men a loincloth.

Footwear for both sexes was made from fabric or soft leather in the form of sandals or boots.

Assyrian and Babylonian hairstyles and headgear

Care of the coiffure was very important for men and women among both the Assyrians and the Babylonians. The hair was grown long and carefully curled and ringleted, with false hair added if needed. Perfumes, oils, and black dye were used on the hair. Men grew long, equally carefully tended curled beards. A band of metal or fabric encircled the brow, or a woolen, felt, or leather cap shaped like a fez was worn. The royal headdress resembled a pleated crown or a mitre and had dependent lappets at the rear. Jeweled ornamentation to the costume was rich and heavy and of high quality.

Reproduced by courtesy of the trustees of the British Museum; photograph: John R. Freeman & Co. Ltd.

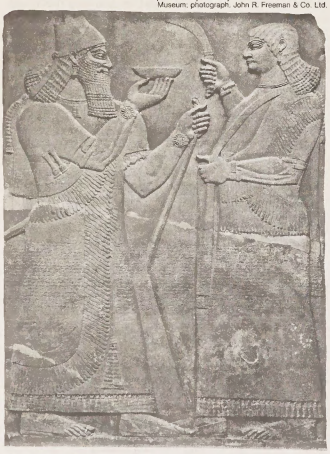


Figure 4: Ashurnasirpal II (left), king of Assyria, with an elaborately dressed beard, wearing sandals and a full-length tunic decorated with embroidery and tassels. Alabaster relief from the palace of Ashurnasirpal II (reigned 883–859 BC) at Nimrūd, Iraq. In the British Museum.

THE AEGEAN: MINOAN AND MYCENAEAN DRESS

The Aegean region and in particular the island of Crete, which was inhabited from about 6000 BC, can be considered the cradle of western European culture. Settlers came to Crete from areas farther east—from Anatolia, North Africa, Syria, and Palestine. By 2500 BC the Cretan civilization was becoming established and, as a maritime people with extensive trade in the Mediterranean and the Middle East, was influenced by many sources. The Cretans created a society and a dress style of their own, one dissimilar from the earlier one of Egypt and the later of Greece. The greatest and most prosperous years were from 1750 to 1400 BC; this was the time of the building of the great palaces, notably Knossos, from where the remains of coloured frescoes, painted vases, and sculpture in marble, terra-cotta, and coloured ceramics have been excavated and are on display in the museums of Iráklion and Athens. Even finer and more complete frescoes have been preserved from the more recent excavations of the Minoan city on the island of Thera (now Santorin), an island largely destroyed in the cataclysmic volcanic eruption of about 1500 BC.

Cretan dress is characterized by its vivid colouring, elegance, and sophistication. It is also notable for the gaiety of feminine attire, typical of a society where women—unlike that of classical Greece—are depicted side-by-side



Figure 5: *Middle Minoan period dress.*
(Left) Woman wearing a bell-shaped skirt with flounces, an open jacket, and a diaphanous vest covering the breasts; her hair is intricately dressed with rows of small curls on the forehead and a chignon from which a ponytail falls. Fresco painting, *c.* 1600 BC. In the palace at Tiryns, Greece. (Right) Priest-king wearing elaborate loincloth attached to a tight, broad belt. Fresco from the palace at Knossos, Crete, destroyed *c.* 1400 BC. In the Archaeological Museum, Iráklion, Crete.
(Left) From *Historia del Arte:* photographs, (left) E.D.I. Studio, Barcelona, Spain, (right) Andre Held, Switzerland

with men, taking part in all the activities of life and not relegated to the domestic background.

Men's garments were few. Chief of these was a loincloth of wool, leather, or linen, tightly belted at the waist and arranged as a short, elaborately decorated skirt. The belt was drawn tight to contrast the slender waist (presuming that the man had one) with the masculine breadth of chest. By 1750 BC women were wearing a long bell-shaped skirt, often in a series of flounces, over a loincloth; with this, they wore a bolero-like jacket that had elbow-length sleeves but was open in front, leaving the breasts bare. In the later period a boned bodice was worn, constricting the upper torso but accentuating the full, bare breasts above. (This is the first recorded example in Europe of corseting constriction of the figure and remained an isolated instance for centuries.)

Women's clothing

The Cretans liked bright colours, and their dress was vividly embroidered and decorated. The hair of both sexes was worn long, looped and braided and dressed with jewels, pearls, and ribbons. The Cretans bathed frequently, oiling their bodies afterward. Men were generally clean-shaven.

Outdoors both sexes wore sandals or shoes. In winter calf-length boots were adopted, and short woolen, fur-lined cloaks were fastened by pins around the shoulders.

With the collapse of the Minoan civilization in Crete about 1400 BC, a new culture arose on the mainland in the Peloponnese, notably in the maritime principalities of Mycenae, Tiryns, and Pylos. As the frescoes from the palace of Tiryns illustrate, the costume was similar but richer still.

ANCIENT GREECE

The long period of Greek culture is customarily classified into three segments. Up to about 500 BC is described as the Archaic period. This was the time when the several different civilizations of mainland and island Greece, Anatolia, and North Africa coexisted, the arts and costume of each influencing the others. The Dorians had invaded the Minoan kingdoms in Crete and the Peloponnese from about 1200 BC. They were a northern race from Illyria and a more primitive society than the Minoans. Modern knowledge of their dress is imperfect, but it seems to have been simple and crude. Woolen cloth, made from the flocks of local sheep, was employed. It was cut into squares of fabric and then pinned on the shoulders and bound around the body. The influence from Anatolia, where the inland climate was more severe, introduced hooded cloaks, banded leg coverings, and Phrygian caps with a point on top.

A later Archaic culture, the Ionian, then established itself in Greece. A more advanced society, the Ionians developed a higher-quality textile industry, producing finer materials in wool and linen that were more suited to a draped style of dress. In the 8th and 7th centuries BC the Ionians developed an extensive trading economy around the Mediterranean region from Gaul in the west to Libya in the east.

The 5th and 4th centuries BC were the years of the great Classical period of Greek culture, the time when a very simple but highly sophisticated and superb quality of work was achieved in the arts, especially in architecture, sculpture, and literature. This was the case with costume as well, the designs of which can be studied in detail from painted vases and sculpture. Classical Greek dress was a draped style, one in which there was little sewing. The garments for men and women were similar, consisting of oblong pieces of fabric in different sizes and materials, draped in various ways and held in place by ribbons and decorative pins. The dress was a totally natural one; there was no constriction and no padding. The simplicity of the dress was offset by the myriad ways of wearing it, a sophistication achieved by personal expression of the wearer.

Classical Greek dress

As time passed and finer materials (mostly linen) were produced, a further variety in draping was created by pleating, a treatment particularly in use for feminine wear.



Figure 6: Woman wearing the Greek chiton and himation. Marble statue from the Nereid Monument, Xanthus, Lycia, Anatolia, *c.* 400 BC. In the British Museum.

Reproduced by courtesy of the Trustees of the British Museum

The pieces of material were set into pleats, soaked in a thin starch solution, twisted and tied at the ends, then left in the sun to dry. This gave a greater permanence to the pleating.

The subject of colour in Greek dress is a difficult one. Neither sculpture nor vases (which are in black, red, and white) provide information. For a long time it was believed that the dress was largely white, and the reintroduction of the "Greek" style in Regency England and Directoire France presumed this from the marble sculpture. It is known, however, that buildings and ornament were painted in bright colours, and so it seems reasonable to believe that in the sunshine of Greece dress was also coloured; literary sources, moreover, report almost all types of colour being employed. In general, tunics were in lighter colours, cloaks darker. Decoration was most often by the classical ornament seen in architecture: the fret (key) pattern, flowers such as honeysuckle (anthemion), circles (paterae), and stripes.

The Hellenistic Age of Greek culture, dating from 323 BC and lasting until Greece became part of the Roman Empire in 30 BC, was a wealthier time, reflecting the wider boundaries of the Greek world resulting from the conquests of Alexander the Great. To the fine linens available in costume were added cotton from India and silk from China; thus the draped mode became more varied and elaborate.

The chiton

From classical times the chief garment was the chiton, a type of tunic made from one or two pieces of material hanging back and front, pinned on one or both shoulders, and girded. For men the chiton was usually knee-length and seamed up one or both sides. An ankle-length version was worn by women and for more formal wear by men. The simplest type of chiton was sleeveless, but later a sleeved version was made possible by using a much wider piece of material pinned at intervals at shoulder level, creating an elbow-length wide sleeve. A variation on the chiton style for both sexes was achieved by wearing a double girdle, one at waist level and one around the hips, the material being bloused out in between.

The peplos was a woman's garment. Made of one or two pieces of fabric, it hung from the shoulder pins to above or below the waist girdle. Alternatively, women used a longer piece of the chiton material and folded it over in front to nang in a similar manner.



Figure 7: Grecian charioteer wearing long chiton. Bronze statue from the Sanctuary of Apollo at Delphi, *c.* 470 BC. In the Archaeological Museum, Delphi, Greece.

Photograph, Toni Schneiders

Giraudon/Art Resource, New York City



Figure 8: Man (left) wearing the himation draped over one shoulder; the two women are dressed in the peplos. Marble figures from a fragment of the east frieze of the Parthenon, Athens, Greece, *c.* 440 BC. In the Louvre, Paris.



Figure 9: Women's dress from the Hellenistic Age, showing the himation draped over the head and covered by a conical straw hat. Terra-cotta figurine from Myrina (near present-day Bergama, Tur.), 4th–3rd centuries BC. In the Louvre, Paris.
Alinari/Art Resource, New York City

There were two chief forms of cloak or wrap. The smaller one—the chlamys—was of dark wool and was worn pinned on one shoulder, usually leaving the right arm free. The larger wrap was the himation, worn by both sexes. Draped in many different ways, it covered the body and could be drawn up over the head. In sculpture, philosophers and statesmen are commonly depicted wearing the himation.

Knowledge of underwear is limited. Literary sources tell of a linen girdle worn to control the female figure and of a band to delineate the breasts. Men wore a loincloth.

Men's hair was long in the early years, but later it was cut short and carefully curled. Bleach was often used to make the hair fashionably blond; perfumes and pomades were applied. Beards were common until the time of Alexander. Most men were bareheaded, a hat being reserved for bad weather. There was a low-crowned, broad-brimmed style—the *petasos*—and a brimless cap, the *pilos*. Women's hair was long; it was usually curled and waved on the forehead and sides and drawn to a chignon at the nape. Many women wore wigs of different shades and decorated their coiffure with flowers, jewels, and fillets. They draped the head with the cloak and, in the Hellenistic period, sometimes perched a straw hat on top.

Footwear

Both sexes went barefoot indoors but outside wore leather sandals. Men also wore boots, which were laced up the front and might be fur-lined.

Greek jewelry was very fine and was, especially in the later centuries, worn in abundance. Both sexes used perfume, and women employed extensive makeup to give brilliance to their eyes, lashes, and cheeks.

ETRURIA

Cultural development came later to Italy than to the Aegean area. The Greeks colonized southern Italy and Sicily from the later 7th century BC, but it was the Etruscans who introduced a high standard of civilization, in the previous century, to the central region of the peninsula. They called themselves the Rasenna (though in Latin they were known as the Etrusci or Tusci). It is believed that they may have emigrated from Anatolia or possibly from farther east. They quickly developed their culture in their new land and, soon after 700 BC, they were living in an urban society capable of a high standard of building and visual arts. In dress, as in the other applied arts, they drew their inspiration and knowledge from a mixture of sources, chiefly Greek and Oriental.

Etruscan society appears to have had more in common with the Minoan culture than with that of Classical Greece. This was true, for example, of the position of women. Unlike the custom in Greece and Rome, where women were relegated to a submissive, domestic role, in Etruria women shared all the activities of life with men. The wealth of pictorial evidence that exists, chiefly the coloured frescoes and sculpture found in the great burial places such as the necropolis at Tarquinia, depicts women taking full part at banquets, dances, and concerts as well as attending racing, athletic, and other types of contests. These sources also indicate a close affinity of dress with the Minoan, illustrating sewn, fitted garments, bright colouring, rich decoration, and an abundance of beautiful jewelry—a craft at which the Etruscans excelled, especially in gold. Nevertheless, Etruscan dress, for both sexes, demonstrates a marriage between East and West, blending Eastern features from Egypt, Syria, and Crete with a later Ionian-style draped attire probably derived from the contemporary Greek colonists in southern Italy. Thus, Etruscans can be seen wearing both draped, pinned tunics and fitted, sewn ones, or such Greek styles as the chlamys, himation, or chiton in conjunction with footwear with Middle Eastern-style turned-up toes. Some Etruscan garments presaged later styles; for example, the *tebenna*, a semicircular mantle, was an early version of the Roman toga, and a decorative collar derived from Egypt anticipated a later Byzantine version.

Etruscan blend of Eastern and Western styles of dress

ANCIENT ROME

The Roman civilization lasted from the traditional founding of the walled city in the mid-8th century BC to the

Scala/Art Resource, New York City



Figure 10: Etruscan musicians wearing tunics, cloaks similar to the Greek chlamys, and sandals. Detail from a fresco in the Tomb of the Leopards, 5th century BC. In the necropolis at Tarquinia, province of Viterbo, Italy.

final collapse of the western part of the empire in AD 476. Until the 3rd century BC the Romans derived their culture from the Greeks and the Etruscans but after this gradually began to develop their own civilization and to expand their influence, taking over territory after territory—first that of the Etruscans, then Sicily, Carthage and North Africa, Greece in 146 BC, and Egypt in 30 BC. They went on to found the great Roman Empire, which by the 2nd century AD extended from Spain to the Black Sea and from Britain to Egypt.

The history of Roman dress is paralleled by that of their arts and architecture. They inherited their style from the Greeks, but, as the empire extended its borders, incorporating peoples of different customs, climate, and religion, a greater complexity became apparent in response to a greater variation of need. The original pattern had become a mutation. In costume, as in art, the trend was toward a more ornate, richly coloured, more varied, and, especially in the later days of the empire, very luxurious attire. Roman dress also reflected a distinct division of social class, with certain colours, fabrics, and styles reserved for citizens and important personages.

With the expansion of the empire and consequent wider trading possibilities came also the availability of more varied and elegant fabrics. Cotton from India and silks from the East were accessible to the wealthy, enriched by high-quality embroidered edging and fringing. Elagabalus (AD 218–222) was the first Roman emperor to wear silk. Later, looms were set up to weave silk, but China retained control of sericulture, exporting only silk thread or fabric, both of which were prohibitively expensive.

The art of dyeing and knowledge of the use of mordants was now more extensive. The famous dye of the classical world was Tyrian purple, so called because its centre of production was in the twin cities of Tyre and Sidon (now in Lebanon). The dye was obtained from small glands in the mollusk *Purpura* and was costly owing to the small size of the source material. Thus, the wearing of the purple was reserved for a few. (Although the name *Purpura* gave rise to the word purple, the colour was actually a crimson.) Under the empire, production sites were established in Crete, Sicily, and Anatolia. At Taranto in southern Italy a hill survives that is composed entirely of the shells of the *Purpura* mollusk.

The toga

The garment for which Rome is most famous is the toga. A large piece of material wrapped around the masculine body as a cloak, the toga served a similar function as the Greek himation, although the fabric was of quite a different shape. Under the empire, the toga acquired a special distinction because of its unique and complex method of draping and because, as a note of rank, its wearing was restricted to Roman citizens. The toga was not rectangular in shape like the himation but was a segment of a circle, measuring about 18 feet along the chord of the segment and about 5.5 feet at its widest point. It was made of wool and so was very heavy. To drape it, about five feet of the straight edge of the fabric was placed against the centre front of the body from ground level upward. The rest of the material was then thrown over the left shoulder and passed around the back, under the right arm, and once again over the left shoulder and arm. The right arm was therefore left free. The material could be pouched in front as well as drawn up over the head. Certain patterns and colours were worn by specific members of society.

The toga gradually became a ceremonial garment, and a great variety of other cloaks were worn by civil or military personnel. Some cloaks were hooded; many were like the Greek chlamys. They had a variety of names: *paenula, abolla, paludamentum, lacerna,* and so on. A less bulky, more manageable alternative to the toga was the *pallium,* a version of the Greek himation.

The basic masculine garment was like the chiton; it was called a *tunica.* Colours differentiated the social classes—white for the upper classes, natural or brown for others. Longer *tunicas* were worn for important occasions. About AD 190 the *dalmatic* was introduced from Dalmatia. This was a looser, ungirded style of tunic with wide sleeves.

Feminine dress was very like the Greek, with the Roman woman's version of the chiton called a *stola.* As time passed, women took to wearing several garments one on top of the other, while the garments themselves were made of finer fabrics and were more lavishly decorated. The feminine cloak, the palla, resembled the Greek himation.

Underwear for both sexes consisted of a loincloth—like

Alinari/Art Resource, New York City



Figure 11: (Left) The Roman toga, worn pouched in the front and drawn up over the head. Marble statue of Caesar Augustus, 1st century AD. In the Museo Nazionale Romano, Rome. (Right) Imperial Roman long-sleeved *tunica.* Statue of Commodus, reigned AD 180–192. In the Vatican Museum, Rome.



Figure 12: The Roman *stola* and palla. Marble statue of Agrippina the Elder, 2nd century AD. In the Capitoline Museums, Rome.
Alinari/Art Resource, New York City

briefs—and women also wore a breastband—the *mamillare*.

Footwear was based upon the Greek but was more varied. Apart from sandals, several styles of shoe and boot existed, once again the colours denoting social status.

Both sexes spent a great deal of time on their toilette, in bathing and using perfume and cosmetics. Face powder, rouge, eye shadow, and eyeliner were lavishly applied by upper-class women, who also attached beauty patches to their faces. Wigs and hair switches were commonly worn, and certain colours of hair were fashionable; for example, during the Gallic and Teutonic campaigns, blonde wigs made from the hair of captured slaves were in vogue.

ANCIENT NONCLASSICAL EUROPE

Animal furs and hides made up the chief garments during the Stone Age. They would be held to the body by a thong belt and by pins at the shoulder. Later such skins were sewn to give a closer fit. Finds from tombs and living sites indicate that the people had a fair knowledge of simple weaving, of sewing, and of how to dress skins. No actual garments have survived, however; remains include boxwood and bone combs, reindeer horn buttons and plaques, and decorative items such as necklaces and armlets of beads, amber, and ivory.

The advent of the Bronze Age varies from one part of Europe to another. The art of bronze working came to Italy from the Middle East and then spread westward to Britain and Scandinavia. During the years 1500–600 BC

© The Board of Trustees of the Victoria and Albert Museum

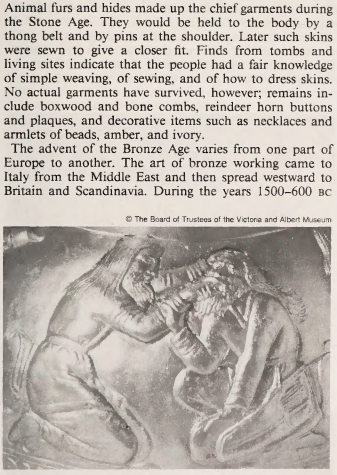


Figure 13: Scythian dress, consisting of tunics, trousers, and short boots. Detail from a replica of an electrum vase (original in the Hermitage, St. Petersburg), from the tomb at Kul Oba, near Kerch, Crimea, Ukraine, 4th century BC. In the Victoria and Albert Museum, London.

the arts of spinning and weaving were further developed; simple natural dyes were used; and decoration was by embroidery, fringing, and plaiting.

Bronze Age garments

In Denmark, northern Holland, and Germany the peat bogs have preserved a number of actual, almost complete, Bronze Age garments; most of the garments, which were found at burial sites, are woolen items that were maintained in remarkable condition in oak log coffins. They include large semicircular cloaks, felt caps, tunics with leather straps and belts, and, for women, jackets and skirts with ornamental belts and hair ornaments. Many of these are on display in the National Museum in Copenhagen and the Schleswig-Holstein Museum of Prehistory and Early History in Germany.

A different type of dress was worn by the nomad peoples who roamed the European and Asian steppes between Manchuria and Hungary. Such groups, which included the Scythians, Cimmerians, Sarmatians, and Parthians, traveled immense distances on horseback, their attire being suited to their way of life. Both sexes wore similar garments consisting of a woolen tunic over a shirt and wide trousers. These garments were worn in layers one on top of another; they were fairly close-fitting but loose enough for comfort and for the practical needs of hours spent on horseback. Short boots were pulled up over the trouser bottoms and tied in place. These peoples also wore leather belts around their waists, and felted woolen caps kept their heads warm. Around 600 BC the Scythians lived in the region around the Black Sea and then gradually moved westward to Romania, Hungary, and Germany. Excavation of their burial sites in the Dnieper valley and near Simferopol, both in Ukraine, and in the Balkans has yielded both actual garments and a wealth of relief sculpture, vases, and plaques that illustrate Scythian dress.

The 6th-century BC Hallstatt culture of the Bavarian and Bohemian areas had an advanced life-style for its time. Finds from this early phase of the Iron Age, however, are chiefly weapons and jewelry. In the 4th century BC the Celtic peoples from central Europe invaded Italy and moved on to Britain, Ireland, and Spain. Finds of the Celtic culture, which consist largely of jewelry, toilet articles, and ornaments, illustrate both the high Celtic standard of craftsmanship, especially in metal, and the individual character of their design. Museums in many countries—notably Italy, Spain, Portugal, Ireland, Britain, and Czechoslovakia—display a wealth of such work.

Roman influence

Roman influence on the dress of the northern and western countries of the empire was strong until the early 5th century AD. This was to a certain extent, however, a two-way influence since, in the colder northern areas, the Romans found the indigenous dress styles of belted tunics with trousers or leg-banding more suitable than their own classical *tunica* and bare legs. Useful evidence of local attire in Britain, Gaul, and Germany is graphically illustrated on the two famous victory columns in Rome—that of Trajan and of Marcus Aurelius. This evidence is reinforced by the written accounts of Roman historians such as Cornelius Tacitus of the 1st century AD and Sidonius Apollinaris of the later years.

THE PRE-COLUMBIAN AMERICAS

Archaeologists now believe that the Americas were first inhabited by people who crossed the Bering Strait from Siberia to Alaska some 25,000 years ago. This narrow strait is shallow and, for a considerable part of the year, frozen over, making such a route easy of access. Over many hundreds of years these peoples gradually spread throughout North and South America, reaching the southerly point of Tierra del Fuego about 9000 BC. At the time of their first encounter with European explorers, this population was composed of societies of many types and levels of achievement, extending from the Eskimo in the north to the North American Indians of the plains, the lakes, and the forests to such highly advanced cultures as the Inca, Maya, and Aztec farther south. The climatic variation over the Americas is immense, presenting almost every type possible in the world. Probably owing to the warmth of the central and southern areas, population density by the 15th century AD was much greater in South America

than in North; the total population is estimated to have been about 20 million in the south and just over 1 million in the north.

The North American Indians. Clearly the costume worn in these different areas varied as much as the climate. It was dictated not only by the weather but also by the materials available for clothing, the standard of culture of the individual groups and tribes, the terrain they inhabited, and the predators they faced. Even the North American Indians alone displayed considerable variety in their dress. There were, however, similarities in dress whether the tribes were of Woodlands or Plains Indians.

Few woven materials were made. The North American Indians wore mainly skins of any of the animals living in their area: deer, elk, buffalo, moose, beaver, otter, wolf, fox, and squirrel. They were highly skilled in tanning such skins by the chamoising process, in which oil is used. The Indians employed animal oils, particularly those found in the brains of the animal, especially deer, so producing a softly textured material that they then dyed in brilliant colours. They made use of the entire skin, adapting the garment to the shape of the animal and wearing it draped and sewn only minimally; the legs, paws, and tail were left attached and hung down as decoration. Two skins were often used for a woman's dress or man's tunic, one back and one front. They pierced the skin with bone awls (not needles with eyes) and threaded edges together with animal sinew. Decoration was by porcupine-quill embroidery, the quills being softened by chewing and then dyed. Garments were also decorated by fringed edging.

Men wore a breechclout and women a short skirt. In warm climates an apron back and front was added to this, along with a cloak or poncho in bad weather. In cooler areas men wore a loose hip-length tunic and thigh-length leggings, the latter tied to the waistband of the breechclout. Women wore a long dress and short leggings.

Hairstyles

Hair was carefully tended by both sexes. For the men there were many, varied styles; in some areas hair was grown long and plaited, in others it was worn loose. Some styles were dramatic, consisting of, for example, a ridge of hair sticking up along the crown of the head, extending from the forehead back, with the remainder of the head shaved. (This style was revived in punk coiffures of more modern times when it was called the Mohawk or *Mohican.*) Animal hair and feathers were added to many hairstyles. The famous feather bonnet headdress, with buffalo horns and headband, was the warbonnet, into which might also be incorporated ermine tails and quillwork. Women's hair was long, worn loose or plaited, and held in place by a headband.

The moccasin was the traditional Indian shoe, a style introduced into Europe in the 20th century. It was made from one piece of soft leather, which enclosed the foot, with no added sole or heel. It was seamed to an inset decorative piece on top of the instep. The leather was then folded over at the back.

Facial and body hair was often plucked out with tweezers, and both face and hair were painted. Red pigment was frequently used to paint the body. Both sexes tattooed their bodies, sometimes all over. Colour was then impregnated into the tattooed skin; bright red was most often used for this.

The Eskimo. The Eskimo settled in Alaska about 1000 BC and gradually spread across the northern coastal region of Canada and as far east as Greenland and Labrador. Their clothing, which was adapted to the Arctic cold, had much in common with that worn in the Siberian Arctic areas from whence they had come. This clothing was, like that of the North American Indians, made from animal skins, but, because of the climate, it was sewn and tailored to the body to keep out the wind. The fur or pelt of the animal was retained and usually worn inside. Thread was of animal sinew and needles of bone or ivory. Like the North American Indians, the Eskimo used all available animal skins: polar bear, deer, caribou reindeer, antelope, dog, fox. They also used birds—the skin for clothing, the feathers as decoration. Sealskin was ideal for boots, the fur turned inward.

Both sexes wore the same type of garments: a hooded tunic or coat, trousers, and boots. The hooded tunic was variously named in different areas. Two of these—the parka of the Aleutian Islands and the anorak of Greenland—have become essential items of modern dress. In winter two or more tunics might be worn; these were known as *kuletak.* Trousers were *kallik.* and boots *kamik.*

Untanned, untreated animal skin suffers from putrefaction and does not last too long. To tan leather, the exterior layer of fur and blood vessels must be removed, as well as the inner layer, leaving the central dermis to be tanned to become leather. Unlike the North American Indians, the Eskimo needed to retain the fur layer, so they scraped off the inner layer and treated the dermis by soaking and manipulation. They soaked it in a liquid of urine and ash. The women, who did this work, chewed the skin to soften it. They also made waterproof capes from the intestines of seals and walrus.

The Aztec, the Maya, and the Inca. At the opposite end of the climatic scale were the cultured peoples of central and southern America who flourished during the European Middle Ages; these were, notably, the Aztec, the Maya, and the Inca. The Aztec and Maya lived in a hot climate and so wore a minimum of clothing, although their garments were brightly coloured and decorated. The Aztec settled in Mexico about the 12th century. Their capital city, Tenochtitlán, which they established in the 14th century, was on the present-day site of Mexico City.

Aztec dress

The men wore loincloths, the women tunics and skirts, all made from cotton. Ornamental cloaks were retained chiefly as garments of rank rather than of necessity. The decoration of Aztec costume was chiefly by exotic plumes, but fur also was used. They wore a great deal of jewelry, mainly of gold.

The Maya came to Guatemala about 800 BC and spread into the Yucatán Peninsula. Their culture, which was then a Stone Age one, flourished chiefly between AD 250 and 900. They also wore few garments: a loincloth for men and a cloak when needed; women wore a loose sleeveless dress or blouse and skirt. Cotton and sisal were cultivated, with the women spinning and weaving these fibres. The Maya also developed a method of tie-dyeing the yarn and of weaving patterns using bright colours for dyes. Embroidered decoration also was practiced.

Cotton fabrics were mainly reserved for upper-class wear, as were beautifully decorated leather belts and sandals. For the ordinary people, tapa—a cloth derived from tree bark as in Polynesia—was made. An important part of

Reproduced by courtesy of the Trustees of the British Museum



Figure 14. Mayan men of the upper class wearing decorated loincloths and ornamental headdresses. Detail from a polychrome vase, Quiché Maya, from Nebaj, Guat., Late Classic Period (6th–10th century). In the British Museum.

Mayan decoration was provided by feathers from the birds of brilliant plumage, which were available and which were skillfully incorporated into the weaving processes. The feathers were also widely used in the ornamentation of headdresses. The long, iridescent tail feathers of the quetzal were especially prized, as they were also in Aztec dress.

The Inca came from the valley of Cuzco in the high mountains of Peru. During the 15th century they established a powerful empire of some five million people in the area of what is now Peru, Bolivia, and Ecuador and extending into parts of Argentina and Chile. They also adopted a brightly coloured attire decorated by feathers—indeed some of their fine cloaks were made entirely of feathers woven into a cotton fabric base. They kept herds of llamas for wool and hunted animals for fur; the chinchilla provided the most beautiful of their pelts.

The actual garments were simple: a basic loincloth for both sexes and, over this, a short tunic for men and an ankle-length dress for women. The poncho was the most usual cloak. People went barefoot or wore sandals. They also frequently went bareheaded or, in bad weather, adopted a woolen cap or turban.

THE MIDDLE EAST FROM THE 6TH CENTURY

The style of costume worn over this large and not very well-defined region has been remarkably constant for centuries. This is partly because it has evolved as one suited to carrying out hard work on the land under considerable extremes of climate, serving as a protection to the body against heat, dust, and blazing sunshine, and partly because the wearing of traditional clothing has been accepted and supported by Muslim countries. The actual garments worn are loose-fitting ones that cover, even envelop, much of the body. The names of these garments vary from country to country, but the similarity is clear. Likewise the materials from which they were, and still are, made vary according to what is available. In general, linen, cotton, and wool are the norm, but the well-to-do have always worn garments made from rich fabrics with a silk base. Several of the most famous of these materials originated in this area, including baldachin, the richly decorated fabric with a warp of gold thread and a weft of silk, named after the city of Baghdad, and damask, named after Damascus (in Syria), from whence this richly patterned silk fabric came.

Middle Eastern fabrics

A number of the traditional garments were originally derived from ancient cultures in the region, particularly from Persia (now Iran) and from farther east in India, Mongolia, and Asian Russia; the caftan is one such example. Still worn in comparatively modern times, this is an open, coatlike garment termed, in ancient Persia, a *candys* or *kandys*. Also worn extensively in the cooler climates of Mongolia and China, the style extended westward to become, eventually, the fashionable dolman of the late Ottoman Empire.

The spread of the characteristic costume of the Middle East was due in large part to the Arabs. These were people of the Arabian Desert who, by the 6th century AD, led a stable, rural life in the border areas of Yemen, Syria, and Iraq but who, in the interior region, were largely Bedouin nomads raising camel herds for a living. By AD 750 the Arab empire extended from Spain and Morocco in the west to the Caspian Sea and the Indus River in the east. The chief garments worn at that time were a loose shirt, chemise, or robe; a draped cloak; wide, baggy trousers; and a headcloth or turban. Remarkably similar versions of these may still be seen on the streets of Cairo, Istanbul, or Damascus.

The simple basic garment for both sexes was a loose, long shirt, chemise, or tunic, which often had long sleeves. Over this men wore a robe or mantle of various types. The *aba* or abaya was of ancient origin and is mentioned in the Bible as the attire of Hebrew prophets. It was traditionally made of heavy cream-coloured wool decorated with brightly coloured stripes or embroidery. A voluminous outer gown still worn throughout the Middle East in the Arab world is the jellaba, known as the *jellabah* in Tunisia, a *jubbeh* in Syria, a *gallibiya* in Egypt, or a *dishdasha* in Algeria. The garment generally has wide, long sleeves, and the long skirt may be slit up the sides; some styles are open in front like a coat or caftan. Outer gowns or cloaks sometimes incorporated head coverings. These included the haik, which was an oblong piece of material (generally striped) that the Arabs used to wrap around their bodies and heads for day or night wear; the material measured about 18 feet by 6 feet. A similar mantle was the burnous, a hooded garment also used for warmth day or night.



Figure 15: Egyptian agricultural worker wearing traditional jellaba, Jordan Valley, Jordan.
© Bill Lyons, 1992

The loose, baggy trousers traditional to the Middle East, as well as to the Balkans and Anatolia, are still widely worn by both sexes. The garment is believed to have originated in Persia, and it is presumed that the Arabs saw it there when they invaded that country in the 7th century. The trousers, called *chalvar, chalwar,* or *şalvar* according to the country where they were worn, measured about three yards across at the waist and were drawn tight by cords. The full, leg portion was tied at each ankle. A broad sash then encircled the waist, on top of the *chalvar.* Worn in this way the garment was ideal for working in the fields because it allowed freedom of movement and protected the lumbar region of the spine, especially while bending, from chills. For centuries the garment has also been adopted by men in the fighting forces. Cotton is the usual material for working attire, but fashionable ladies wear such a *chalvar* made from a brocade or silk fabric over linen drawers.

The *chalvar*

The tradition for women to cover themselves from head to toe and veil their faces when they go out in public is an old one, predating Islām in Persia, Syria, and Anatolia. The Qur'ān provides instructions giving guidance on this matter but not a strict ruling. It has been the rigidly male-dominated world of the Middle East that has insisted on the strict veiling of women in public. The enveloping cloaks worn by women for this purpose are similar to one another and often incorporate a mesh panel through which women may peer at the world outside. The most common names for this garment are chador, *chādar, chadri, çarşaf,* and *tcharchaf.*

The characteristic masculine Arab headdress has been the kaffiyeh. It is still worn today, although it may now be accompanied by a Western suit. Basically, the kaffiyeh is a square of cotton, linen, wool, or silk, either plain or patterned, that is folded into a triangle and placed upon the head so that one point falls on to each shoulder and the third down the back. It is held in place on the head by the agal (*igal, egal*), a corded band decorated with beads or metallic threads.

Footwear was in the form of sandals, shoes, or boots, with the toes slightly turned up. Women traditionally wore decorative wooden pattens called *kub-kobs* to walk about in muddy unpaved streets.

THE BYZANTINE EMPIRE

In AD 324 the Roman emperor Constantine decided to rebuild the great city at Byzantium, then a Greek centre, sited strategically on the Bosporus, whose narrow waters connecting the Mediterranean and the Black Sea acted as a gateway between West and East. Constantine called his city New Rome; it was later renamed Constantinople (now Istanbul) in his honour. After the collapse of the western part of the Roman Empire, which was based in Rome (and later Ravenna), in the 5th century, Constantinople became the capital city of a Christian-dominated empire whose extent fluctuated considerably until its own collapse to the Turks in 1453.

Owing to the site of its capital city, the Byzantine Empire was subject to a complex of influences that were nowhere more marked than in the dress of its ruling classes. Over the centuries there were two notable periods of wealth and prosperity, and these, as always, were reflected in costume. The first period was in the time of the emperor Justinian, who reigned from AD 527 to 565. Until this time the influence of Rome was still strong, and dress styles tended to be draped in the fashion of the later years of the empire. There were differences, however, derived in part from Persian and Anatolian designs, that introduced more Eastern forms of dress, such as sewn, closer-fitting garments and richer ornamentation and jewelry.

Because of its success as a trade centre between East and West under Justinian, the Byzantine Empire became extremely wealthy. This wealth, allied to the love of richness and to the high standard of Arabic and Oriental craftsmanship available, led to a costume of magnificent splendour that became the envy of the known world. Luxury fabrics from Asia, Syria, and Egypt became available in quantity and were utilized, despite the high cost, by the leading members of society.

Introduction of sericulture

The domestic textile industry was also stimulated, its development greatly aided by the introduction of sericulture into Constantinople. The Chinese had guarded secrets of the manufacture of silk for hundreds of years, but by about 1000 BC silkworm culture and silk manufacture had been established in northern India, and the knowledge later percolated through to Korea, Japan, Persia, and Central Asia. Justinian had tried early in his reign to divert the silk trade from its route from Persia but without success. He was then presented with a tremendous opportunity when two Persian monks, who had worked as missionaries in China and had studied the process of sericulture and the weaving of the filaments, agreed to smuggle this knowledge, as well as the necessary silkworm eggs, to Constantinople in exchange for a large monetary reward. Silkworms flourished in Constantinople, and the authorities there, like the Chinese and others before them, guarded the secrets of the process and controlled their monopoly in Europe until, inevitably, in the early Middle Ages, the knowledge and means were once more disseminated, this time to Anatolia and Sicily and from there gradually to Italy and France.

During Justinian's reign, Byzantine textile manufacturers produced beautiful, glamorous fabrics, mainly of silk interwoven with gold and silver metallic threads and interspersed with pearls and jeweled embroideries. The use of these heavy lustrous fabrics gradually altered the style of dress; the stiff, ornate materials lent themselves to a simpler cut with only a few folds to break up the often allover, large-motif designs. For both men and women a more fitted, sewn tunic, cinched at the waist by a richly decorated wide belt and hanging straight to knee or ankle, replaced the Roman draped *tunica*. A rich, deep decorative collar, like the preceding Egyptian and Etruscan versions, covered the shoulders. The influence of the Christian church could be seen in the fact that the limbs were generally covered by long, usually fitted sleeves and cloth or silken hose. Cloaks, pinned at the shoulder, were worn outdoors. Imperial dress was characterized by the extensive use of purple and gold. Garments for the wealthy were vividly coloured in reds, yellows, and greens. Such attire is depicted in the 6th-century mosaics and the Ravenna churches of San Vitale and Sant'Apollinare Nuovo.

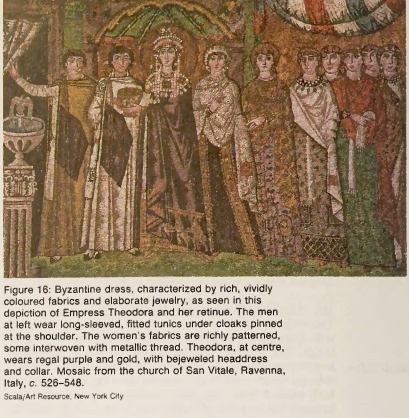


Figure 16: Byzantine dress, characterized by rich, vividly coloured fabrics and elaborate jewelry, as seen in this depiction of Empress Theodora and her retinue. The men at left wear long-sleeved, fitted tunics under cloaks pinned at the shoulder. The women's fabrics are richly patterned, some interwoven with metallic thread. Theodora, at centre, wears regal purple and gold, with bejeweled headdress and collar. Mosaic from the church of San Vitale, Ravenna, Italy, *c.* 526–548.
Scala/Art Resource, New York City

The second period of expansion and prosperity came between the 9th and 11th centuries. Court dress became richer than ever, encrusted with jeweled embroideries and dyed in deep colours, especially purples and reds. Imperial dress included a long panel of gold-embroidered material, which was wrapped around the body with the end hung over one arm. The classical line had completely given place to an Eastern form of dress. For example, the caftan had been adopted as formal wear. Open down the centre front, this coatlike garment was shaped to fit at the back. For both sexes the caftan was accompanied by trousers, not full like the Middle Eastern *chalvar* but more elegantly and closely cut, especially on the lower limbs where they were tucked into boot tops or worn over shoes. The caftan

Byzantine dress strongly influenced that of eastern Europe, especially the Balkans and Russia. Some of the bejeweled silk formal garments were gradually adopted by the church to become vestments in the Middle Ages.

Men tended to prefer leather boots in footwear, black for normal use and red at court. Women sometimes wore sandals but more often were found in soft, ankle-height shoes, brightly coloured and embroidered.

Masculine hairstyles were short, and men were mostly clean-shaven. Outdoors they adopted the Phrygian cap or a hood. Ladies wore veils and often encased their long hair in a silk cap or a pearl net. Elaborate jewelry was characteristic of all Byzantine dress of the upper classes. Perfume was liberally applied but cosmetics less so.

MEDIEVAL EUROPE

The dress of Europeans during the years from the collapse of the western part of the Roman Empire in the 5th century to about 1340 was slow to change and was largely standardized over a wide area. Regional variations in climate, social customs, and relative wealth brought only a limited differentiation in dress.

Clothes for men and women were similar, being sewn albeit crudely and loosely cut. A shirt or chemise and braies—that is, a roughly fitting kind of drawers—constituted underwear. These were of a natural coloured linen. The shirt was hip-length for men, longer for women. It had a round neck, slit in front for ease of donning, and was tied with a drawstring; the braies were similarly fastened at the waist. On top of this was worn one or more tunics—knee- or ankle-length for men and ground-length

for women. The tunic had a round neckline and long sleeves cut in one with the garment; it was loose fitting but girded at the waist. Tunics were made from coloured linen or wool and were decorated with embroidered bands at the neck, wrists, and hem. Legs were covered with ill-fitting hose, these were cut from cloth in two vertical sections and sewn together. They were held up by banding or garters.

Thirteenth-century dress was noted for its plainness. There was little or no decoration, and garments were unbelted. A sleeveless surcoat was generally worn over the tunic. This had derived in the late 12th century from the tabard, a garment worn by crusading knights over their armour to prevent the sun from reflecting off the metal and making them visible to an enemy. The surcoat, which was worn by both men and women, often had slits (called fitchets) on each hip so that the waist belt underneath with purse attached could be reached without fear of thieves.

Men's hair might be long or short; some men were clean-shaven, while others had beards. Ladies wore their hair long, parted in the centre, and plaited and then pinned up at the sides; they then pinned a white linen neckcloth to the plaits on each side (the wimple), concealing the hair, and on top of this wore a veil, a white linen crown, or a pillbox cap. Such headdresses were known by a variety of names, including barbette, fillet, and *touret*.

Introduction of tailored garments

Toward 1350 a great change occurred in costume. Clothes were tailored to fit and display the human figure instead of obscuring it. As with most fundamental events in human history, this had more than one underlying cause. It was due partly to the gradual improvement in the ability to tailor garments; partly to the availability of a higher quality and greater variety in fabrics, which were now reaching the West from Italy and farther east; and partly to society's readiness to receive such developments. The most important reason was the spread of the Renaissance movement from Italy. A movement both spiritual and secular, the Renaissance was dedicated to reviving classical concepts and to celebrating the dignity and importance of man. This was expressed in costume by the beautification and display of the human figure.



Archives Photographiques, Paris

Figure 17: *Typical simplicity of 13th-century European dress.* Man (left) wearing a surcoat with hanging sleeves and a slit skirt showing fur lining; the woman wears a loose surcoat that, like the man's, reveals the sleeves of the garment underneath. Statues on the right-hand portal of the west facade of Strasbourg cathedral in France, *c.* 1280–1300.

During the remainder of the medieval period, men were the more gloriously appareled of the sexes, and so it was men whose bodies were more specifically delineated. In the 14th century this concept was initiated by the establishment of the fitted tunic, which was cut into four sections that were seamed at the centre back and at the sides and fastened with buttons centre front. By 1340–45 this tunic was hip-length with a heavy leather belt decorated with metal and jeweled brooches encircling the hips only a few inches above the hem. Sleeves were elbow-length, ending in a cuff from which a two- to three-foot streamer, called a tippet, depended. The undertunic, of similar cut, had long sleeves, buttoned to fit closely from elbow to wrist.

The hose were now fitted more closely also. They were cut from velvet, silk, or woolen cloth in four sections and extended from the foot to the upper thigh, where they were attached by points (laces with metal tag ends) to the lower edge of the undertunic. By 1370–80 the hose grew longer to become tights and were laced by points all around the body to the by-then waist-length undertunic. As outer tunics also became increasingly short in the early 15th century, a codpiece became necessary. This was a bag covering the front opening between the two legs and was attached by points to the hose. (The name derives from the medieval term cod, meaning bag.)

Ladies' gowns also were affected by current events. The neckline was lowered and was cut straight across at shoulder level. The bodice, which extended to the hips, was fitted like the men's tunic, and a similar heavy belt encircled the hips. Below the hips the skirt was gored, very full, and long. Sleeves resembled men's styles. Another gown, called a sideless surcoat, was often worn on top. This had no sleeves but had a very large armhole to display the gown beneath; the armholes and a front panel known as the plastron were often trimmed with fur.

There were several new forms of decoration at this time. One was parti-colouring, in which all garments, including hose, could be of one colour down one side and a different hue on the other, the dividing line thus delineating the form of the figure. Counterchange designs—heraldic, floral, or geometric in motif—were introduced where the ground colour and design colour were interchanged. Edges of garments were cut into various shapes; these were called dagges (Middle English: dags).

Parti-colouring

During the 15th century these trends developed further. Men's hose became still better-fitting. Tunics were shorter, often only waist-length. Fabrics were richer and beautifully patterned. For older men, for whom displaying the figure was less suitable, a long gown was introduced to wear over the tunic. At first (14th century) full and long like a dressing gown (the houppelande), the gown gradually became more tailored and formal, with vertical pleats in back and front. All garments, for both sexes, were fur-edged and, often, fur-lined—for both warmth and appearance.

By the 15th century, styles, accessories, decoration, and fabrics were beginning to vary from area to area. The fashion-setter in the years 1430–75 was Burgundy, a duchy which then controlled Flanders and much of modern France. It was the wealthiest region, and the fabrics it manufactured—velvets, silks, gold and silver materials, and embroideries—were of the highest quality. After Burgundy's defeat in 1477 at Nancy, Italy became the fashion centre of Europe. Italian fabrics were equally beautiful but less heavy and with less fur. Colours were gay and bright, and the emphasis for both sexes was on an elegant, natural human form with a gracious ease of movement.

Men's hairstyles varied greatly during this long period. In general, they were short until the later 15th century, and men were mostly clean-shaven. The main head covering was the hood with an attached shoulder cape and a long extended point, or tail, known as a liripipe. By the 1420s a new way of wearing this hood was tried. The face portion was placed on the head, then the cape was arranged in folds like a cockscomb and tied into place with the liripipe, the end of which trailed over the shoulder (a style immortalized in jester's attire). This was an inconvenient arrangement and so a padded roll evolved—the

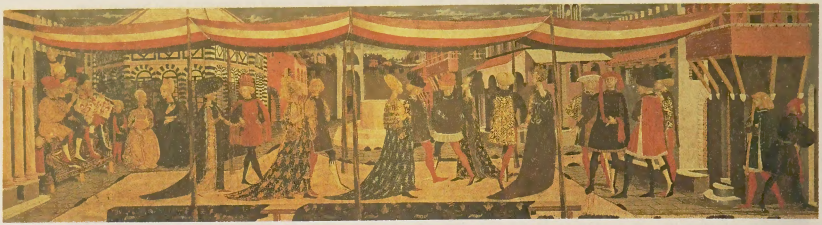


Figure 18: Fashionable mid-15th-century Italian dress. The men wear pleated, fur-trimmed tunics, fitted hose, and, on their heads, the roundlet and liripipe. The women are dressed in long gowns made from richly embroidered fabrics and a typical selection of the varied styles of headdress of the time, from turbans to heart-shaped designs. Detail from the "Adimari Wedding" cassone, Florentine, *c.* 1470. In the Accademia, Florence.
Scala/Art Resource, New York City

roundlet—with the separate shoulder cape sewn in place to one side and the liripipe to the other. Toward the end of the century, various styles of tall or broad-brimmed hats, decorated by coloured plumes, replaced the hood.

Women's headgear

Ladies' headdresses were extremely varied. Hair was still long, plaited, and coiled over the ears. These coils might be enclosed in metal mesh jeweled nets called cauls and were worn with a veil. In the 15th century turbans were fashionable—Byzantine style introduced in Italy—as were also steeple headdresses resembling dunce caps and, alternatively, shorter fezlike caps. All were made of rich fabrics and accompanied by veils, either in a soft flowing mode or formed into winglike shapes by wire framework underneath.

Footwear was similar for both sexes. Hose might be soled for indoor wear. Outdoors shoes could be worn with wood and cork pattens strapped on to keep the elegant fabrics out of the mud of the streets. Men wore boots for traveling. Long toes were fashionable in the late 14th century, the ends being padded to keep the shape.

The Middle Ages was not a time for heavy use of cosmetics or perfumes, although knowledge of the making of perfume had been earlier introduced into Venice from the East.

EUROPE, 1500–1800

The 16th century witnessed important changes occurring in Europe. The limitations bounding medieval society were gradually being breached; the concepts of the Renaissance were being accepted farther west, in France, Flanders, and, finally, England and Spain. People expected a higher standard of living, and there was an expanding middle class. Europe was also looking outward. From Portugal, Spain, and Italy especially, sailors were voyaging to explore both east and west. Their journeys brought the acquisition of riches, new materials, and precious metals. Costume, as always, reflected all this.

The chief centres of wealth were the pacesetters in fashion. Until about 1510 the style was generated from Italy. After this the Germans and Flemish set the pattern, but from about mid-century it was Spain that dominated the scene.

Styles of the first two decades were a development and expansion of the Italian modes of the late 15th century. Young men wore white silk shirts, frilled and embroidered at the neck and wrists. Over this they wore an abbreviated tunic and close-fitting hose, which were often striped to delineate the masculine limbs. Older men covered the tunic and hose with a long gown, open down the centre front, the edges turned back to display the contrasting lining. Men's hairstyles were long and flowing. Their hats, which were set at a jaunty angle, were made of black velvet and decorated with brooches and plumes. Ladies' gowns had square necklines and were cut low enough to reveal the frilled chemise worn underneath. Sleeves were wide and full, and skirts were held or pinned up to display the undergown.

From about 1520 to 1545 the fashionable shape was governed by the strange mode for padded puffs decoratively slashed. This idea is thought to have been derived from the dress of Swiss and Bavarian mercenaries. Each garment was slashed to show the contrasting colour of the material of the one beneath. Whereas the humanist concept of the Renaissance had led to figure display and elegance, the new modes were influenced by the Reformation of northern Europe, giving rise to darker colours, heavier materials, and bulky garments padded to conceal the figure.

Slashing

The masculine tunic—now called a doublet—had a knee-length, gored skirt, which was open in front to display the now padded protruberant codpiece. Over this was worn a rich velvet gown with fur collar and padded sleeves. Shoes and boots had broad toes and, like all other garments, were decoratively slashed. Short hair styles, small beards, and flat velvet caps worn at an angle were fashionable.

The feminine figure was artificially controlled, not by heavy padding but by a tight underbodice with metal or whalebone strips in the seams to give a small waist

The Bridgeman Art Library/National Museums and Galleries on Merseyside (Walker Art Gallery, Liverpool)



Figure 19: Men's dress of the 16th century, featuring slashed and bejeweled doublet with knee-length, gored skirt; codpiece; fur-trimmed, velvet gown; broad-toed shoes; and flat velvet cap. "Henry VIII," oil on panel by the studio of Hans Holbein the Younger, after 1537. In the Walker Art Gallery, Liverpool.

and slender torso. In contrast, the skirt was shaped by a hooped petticoat made from canvas inset at intervals with circular hoops of wicker to give a cone-shaped silhouette. This fashion had originated during the previous century in Castile, Spain, and by 1500 it had become high fashion there. The Spanish skirt, called a *verdugado*, was bell-shaped, however. About 1530 the cone-shaped hoop was introduced into France, where it was popularized by the queen and called a *vertugade*. The style soon appeared in England, where it was known as a farthingale.

The fashionable lady's headdress was a hood made of dark velvet, with long flaps or folds hanging down the back and sides. The face was framed in front by a jeweled metal frame shaped like a pyramid (the English hood) or a horseshoe (the French hood). Under this was worn a decorative cap which almost concealed the hair.

The costume worn from mid-century until about 1620 was the richest ever seen in the history of European dress. It was made from beautiful fabrics heavily encrusted with embroidery, pearls, and jewels. Fine lawns and lace were employed, and all garments were extensively patterned. During these years Spain was extensively wealthy from the results of the New World exploration, and Spanish dress—which was elegant and tasteful, formal and restrictive, and doubtlessly uncomfortable to wear—was paramount. Paradoxically, when other nations adopted Spanish modes they mostly took them to excess, the Spaniards themselves remaining restrained in their dignified black garments.

The masculine doublet was fitted to the waist and buttoned centre front. Its skirt had now been replaced by trunk hose, which were loose mid-thigh-length breeches gathered into a tight waist and thigh bands; decoration was by embroidered strips called panes. Embroidered clocks decorated the now knitted silk stockings. Shoes had returned to the natural foot form. The dashing Spanish cape had replaced the cumbersome gown. These capes displayed great variety in size, shape, and method of wearing.

Ladies' fashions became more constricted and elaborate as time passed. The boned bodice evolved into a restrictive corset. The farthingale became wider and, by the 1580s, was extended by a padded sausage known as a bum roll or barrel, which was tied around the waist under the skirt. Later the French introduced the wheel farthingale, which was drum-shaped with radiating spokes on top. The gown neckline became very décolleté, almost displaying the breasts. From the 1570s to the 1770s a stomacher—a stiff, V- or U-shaped panel heavily decorated with jewels and embroidery—was often worn over the centre front bodice of the gown.

A characteristic feature of dress of this time for both sexes was the ruff, introduced from Spain. Called a band (ruffs laundered and ready to wear were kept in band boxes), it was a strip of material tied around the neck. Another, ruched strip was sewn on to it. After 1565, with the introduction of starch, ruffs became larger and were often edged with embroidery and lace. The very large "cartwheel" ruffs were not worn in Spain, nor was the wheel farthingale. It was in Holland, Germany, France, and England that the extremes of these fashions, which lasted until about 1620, were seen. The ruff

The Dutch led the fashion world between about 1620 and 1645, setting the mode for a freer, more comfortable attire. With whalebone constriction and distortion by padding removed, the emphasis was once more upon the natural lines of the human figure. By the 1620s Holland was emerging from Spanish control, extending its trade dramatically to become wealthy and influential. The garments worn by the well-to-do were still made from beautiful fabrics, but these now included fine wools as well as velvets and silks. The encrustation of jeweled embroideries was abandoned in favour of a decoration reserved for only certain parts of the costume, permitting the beauty of the fabrics to be appreciated. The material that above all was characteristic of these years was lace, seen especially in the falling bands—large collars covering the shoulders, which had replaced the 16th-century ruffs—and their elegant matching cuffs.

The years 1630–50 have been aptly dubbed by some costume historians as the time of "long locks, lace, and leather." Men grew their hair long and wore it, beautifully cared for, falling naturally onto the shoulders and down the back. Complementary to this coiffure was a large beaver, felt, or velvet hat, dramatically ornamented by coloured ostrich plumes. The leather refers to the fact that the fashionable footwear was a boot rather than a shoe. These boots were made of soft leather; they had heels with platform soles and immense bucket tops, over the edge of which frothed lace-edged boot hose. The doublet had become an elegant hip-length jacket, and the trunk hose were replaced by knee-length breeches tied with a ribbon sash at the knee. Ladies' gowns were simple and stylish; the full skirt fell naturally from a high waistline

By courtesy of Mr. Simon Wingfield Digby, Sherborne Castle, England



Figure 20: English dress of the 1590s. Queen Elizabeth wearing an elaborately patterned dress over a corset and wheel farthingale; a lace-edged open ruff and ropes of pearls adorn the gown. The woman on the right is similarly clad. The men wear doublets, trunk hose, embroidered capes, and ruffs. "The Procession of Queen Elizabeth I," oil painting attributed to Robert Peake the Elder, *c.* 1600. In the collection of Simon Wingfield Digby, Sherborne Castle, England.



Figure 21: *Mid-17th-century Dutch dress.*
Men and woman wearing wide collars falling over the shoulder. Man (right) wearing rhinegraves and the woman (centre) a skirt bunched up to reveal a petticoat. "A Game of Skittles," oil on canvas ascribed to Pieter de Hooch, 1660–68. In the St. Louis Art Museum.
By courtesy of the St. Louis Art Museum, Missouri

and the shoulders were covered, as were the men's by a lace falling band. The hair was dressed high on the crown in a bun decorated with pearl ropes and with ringlets at the sides and brow.

French dominance of European fashion

The *grand règne* of Louis XIV of France lasted from 1643 to 1715. In this time the king established France as a great European power, and from about 1660 France became the unchallenged leader of the European mode, a position it held until almost 1939. The fashions were set in Paris, and knowledge of these styles were disseminated by the mannequin dolls sent out to European capitals and by the costume plates drawn by notable artists from Albrecht Dürer to Wenzel Hollar.

In men's dress the mid-century years represented a transitional period when ribbon and lace ornamentation dominated the whole attire, which consisted of a white shirt, an open, waist-length jacket, and full breeches that resembled a skirt. These breeches were known as petticoat breeches (owing to the effeminacy of their appearance) or, more correctly, as rhinegraves.

Between 1665 and 1670 came a quite different masculine style and one that presaged the traditional three-piece suit of modern times. Initiated in France, this began as a knee-length coat called a justaucorps, the idea deriving from the Persian caftan. It had no collar and was worn open in front. The short sleeves ended in cuffs. By 1680 the sleeves were longer, and under the coat was worn a slightly shorter waistcoat together with close-fitting knee-breeches. At the neck the falling band had been succeeded by an elegant, lace-edged cravat.

Ladies' styles changed less noticeably at this time. The gown neckline was lowered, and the falling band was replaced by a decorative neckband. The waistline was also lowered, and a stiffened bodice was reintroduced to slenderize it. Skirts were fuller and longer but were draped up on each side and fastened with ribbon bows to display the petticoat underneath.

In the last decade of the century both sexes wore a high coiffure. In the case of the men it was a wig. The periwig had been fashionable since about 1670. It was made of naturally coloured hair—human where possible—and consisted of a great curtain of curls and ringlets cascading over shoulders and back, while above the brow the curls rose high on either side of the centre parting. With these full-bottomed wigs the hat, now a three-cornered tricorne, was usually carried under the arm. Ladies wore a tall headdress—the fontange—consisting of tiers of wired lace decorated by ribbons and lappets.

The periwig

Until the early 1770s, French control of fashion was complete. It was in France where the trades and professions vital to fashion were established: dressmaking, tailoring, wig making, haberdashery, millinery. Textiles for these crafts were varied and luxurious. They were beautiful but, unlike their 16th-century counterparts, were painted, embroidered, or printed with dainty rather than large-motif designs and were decorated not with jewels but with lace ruffles, ruching, and ribbon bows. Silks, satins, taffetas, and velvets were preferred until the last three decades of the 18th century when—as a consequence of the infamous "triangular trade" of manufactured goods, slaves, and raw cotton carried on by Europeans, Africans, and Americans—fine cottons became readily available.

Fundamental changes were taking place in society during the 18th century. Man, for hundreds of years the peacock of fashion, now gradually ceded his position; his garments became less ostentatious while women's dress became the vehicle for change and display. This development was partly the result of the increasing importance of women in society. The middle classes were also increasing in numbers and influence at this time, leading to a wave of egalitarianism in dress and a gradual end to the idea that richness and high fashion were the prerogative of the aristocracy.

Reproduced by permission of the Trustees of the Wallace Collection, London, photograph, J.R. Freeman & Co. Ltd.

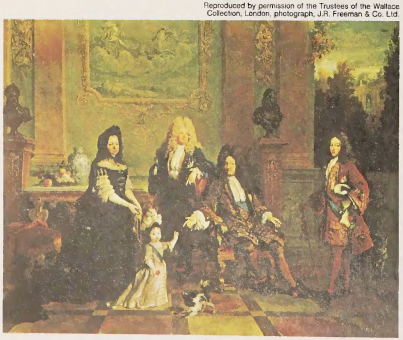


Figure 22: *French dress of the Louis XIV period.*
Male attire of long coat with wide, turned-back sleeves, waistcoat, lace cravat, tight-fitting breeches, and periwig. "Louis XIV and His Family," oil painting by Nicolas de Largillière, 1711. In the Wallace Collection, London.

During the 18th century men continued to dress elegantly, and changes in their costume style were gradual and limited. The *habit à la française,* the French term for the suit consisting of coat, waistcoat, and knee breeches, had become accepted wear. There was a trend away from brightly coloured satins and velvets, however, toward darker, more sombre cloth materials. The cut of the *habit* also became subdued; there was less decoration, and the style fitted the figure more closely. Wigs were worn through the 1780s, in many and varied styles, but became smaller and less elaborate as time passed; powder was used for much of this time. The tricorne hat remained the style of this century.

For ladies there was a return, by 1700, to a rigid corset to slenderize the waist and a framework petticoat to define the shape of the skirt. In the early decades this was a hoop skirt, circular in section and very full. A popular style of gown worn over this was the sack (*sacque*), which had been derived from the informal house dress of the early years of the century. In France this style was often called the *robe volante.* From a low, wide neckline the gown flared out freely over the hoop petticoat. By 1720–25 the

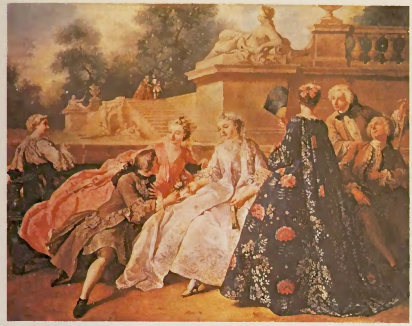


Figure 23: *French dress of the early 18th century.* Women wearing *robes à la française;* men wearing powdered wigs, *habits à la française,* and stockings. "The Declaration of Love," oil painting by J.F. de Troy, 1731. In the Staatliche Schlösser und Gärten, Berlin.

Reproduced by courtesy of the Staatliche Schlösser und Gärten, Berlin

fullness was concentrated at the back in two deep box pleats sewn to the neckband, while the gown was waisted at the front. This was the *robe à la française.*

The panier

Toward mid-century the hoop framework gradually changed shape to become oval. Then known as a panier (French term for basket), it consisted of a basket form on each hip tied in at the waist by tapes. Soon the frame became so wide that women found it difficult to negotiate a doorway or a sedan chair, so a collapsible folding panier was devised, made only of whalebone hoops connected together by tapes. The years 1750–75 saw the most elaborate and outrageously decorated panier gowns, a riot of ruffles and flounces and ribbon bows. It was also the time of ridiculously high, overdecorated, and powdered wigs. Cosmetics of all forms, many containing white lead, mercury, and other injurious chemicals, were copiously used, a reintroduction of the 16th-century practice.

The Bettmann Archive

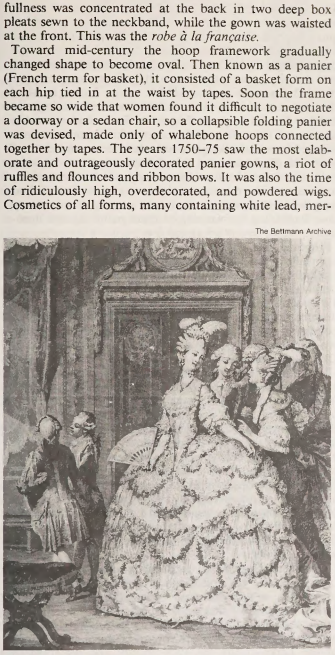


Figure 24: French woman wearing lavishly decorated panier gown and elaborate, powdered hairstyle. "The Queen's Lady-in-Waiting," engraving by Jean-Michel Moreau le Jeune, 1776.

By the 1770s a reaction to this excess was setting in in England, where simpler gowns with a framework petticoat were worn, and the fullness of the skirt was drawn to the back. A sash encircled a high waistline, and a soft fichu was draped around the neck. These gentler yet elegantly feminine styles gradually spread throughout Europe and were finally accepted in France.

Children's clothing

For centuries children had been dressed as miniature adults, but in the 1770s there was a marked divergence from this established custom. Children began to be dressed in more comfortable garments suited to their age. Girls' dresses were rather like the easier styles of their mothers at this time, but boys were dressed in a frilled shirt and ankle-length trousers, the waistband of which was buttoned to the shirt. This costume, in which the wearing of trousers as fashionable dress antedated its introduction for adults by a generation, was oddly entitled a skeleton suit.

(D.Y.)

COLONIAL AMERICA

Seventeenth-century North America was colonized by settlers from northern and western Europe. These settlers brought with them habits and ideas in dress that were characteristic of their places of origin, but their clothes were also influenced by the climate of the part of the country to which they had come. For example, the earliest settlers, the Spanish, arrived in Florida in 1565. There, as well as in their later settlements in Texas and California, the climate was not very different from that of Spain, so that the colonists could continue wearing Spanish styles. In contrast, colonists farther north in New England experienced harsher winters than they had been accustomed to and so found a greater need than they had in England to wear furs and skins.

In all areas many colonists thought it important to preserve class distinctions. Because of this, they passed many sumptuary laws to maintain differentials. Such protocol in dress was a visible expression of their determination to maintain their heritage. Similar laws restricting dress were also passed for religious reasons, reflecting the mid-century Civil Wars in England. In America, as in England, plain dress and rich dress became, in effect, the respective symbols of Puritan and Cavalier. Many Virginia colonists leaned toward the Cavalier; Puritan ideas prevailed in Massachusetts. Virginia dress, though it differed little in design from that of New England, was in general more costly. The Puritans omitted such extravagances as fine brocades, rich laces, ribbons, and feathers; the change to simpler dress that had begun before their departure for America continued.

Difference between American and European clothing

Probably the greatest change in clothing in America, as opposed to Europe, took place in everyday working costume, the Americans wearing heavier and warmer clothing made of stronger and stouter materials. The distinguishing characteristic of all clothes listed in the inventories of the colonization companies is their wearing quality, and the terms "heavy cloth" and "strong durable stuff" are often encountered. Men and boys wore comfortable, durable jackets and breeches, for example, made from deerskin and buckskin tanned to the consistency of fine chamois with the use of animal brains, a process the colonists had learned from the Indians.

For many English colonists the early years were hard. Most people made their own clothes, cultivating flax and cotton and raising sheep for wool. Clothes for everyday wear were plainer versions of those worn back in England. Best clothes were kept for Sundays and holidays; such garments lasted a long time, and most colonists were therefore wearing styles considered old-fashioned in England. For example, men wore breeches full at the waist, a doublet and jerkin, and a hip-length, loose overgarment that had been fashionable in Europe in the later 16th century. This was the mandilion, derived from the medieval tabard. It was now a loose jacket with free-hanging sleeves. It had been adopted by the Puritans, who took it to New England. The colonist version was generally lined with cotton and fastened with hooks and eyes. By mid-century



Figure 25: A woman and child in stylish 17th-century American Colonial dress. Both dresses are made of fine fabrics and have virago sleeves. The child wears a simple linen falling band collar, while the mother wears a more elaborate lace version. "Mrs. Elizabeth Freake and Baby Mary," oil on canvas by an unknown artist, c. 1671–74. In the Worcester Art Museum, Worcester, Mass., U.S.
Worcester Art Museum, Worcester, Mass.

the buff coat had also become a staple garment among colonists in this area. Originally a military coat made of hide, it was durable and warm; it was cut simply in four sections, with or without sleeves.

The everyday dress of women was a short gown of durable material, with a full skirt over a homespun petticoat, covered by a long apron of white linen. The more stylish dress was longer and made of finer material. It often had the virago sleeve—full at elbow and shoulder and drawn in at intervals by strings of narrow ribbon—that appears in most 17th-century portraits of American women and children.

Slashed clothing was fashionable, as in Europe; in the openings made by the slashes could be seen rich materials. In 1634, however, the general court of Massachusetts forbade men and women to make or buy clothes with more than one slash in each sleeve and another in the back. The starched ruff of the early 17th century gave way to the falling band, the common form of which was a broad, plain, linen collar. Both men and women wore this collar and plain linen turnback cuffs.

Stockings were either knitted or cut from woven cloth and sewn to fit the leg. They were attached to men's breeches by points, or strings, which were also used to secure other garments; later, sashlike garters replaced points. Both men and women wore stout leather shoes with medium heels. Men also wore French falls, a buff leather boot with a high top wide enough to be crushed down. After 1660 the jackboot, a shiny black leather boot large enough to pull over shoe or slipper, replaced the French falls; oxfords of black leather were worn by schoolchildren.

Headgear

Both men and women wore a steeple hat of felt or the more expensive beaver. Men also wore the montero cap, which had a flap that could be turned down, and the Monmouth cap, a kind of stocking cap. Women of all ages wore a French hood, especially in winter, when it was made of heavy cloth or fur-lined; this hood, tied loosely under the chin, is seen in many portraits of the time. Sometimes the steeple hat was worn on top of the hood.

Dutch settlements, including New Netherland and New Amsterdam (later New York), were founded in the 1620s. The settlers in these areas were industrious and tolerant, mixing harmoniously with colonists from other nations. They created a wealthy community but placed no restrictions on dress for sumptuary or religious reasons. Their attire was, as it had been in Holland, of high quality and fashionable but not ostentatious.

French colonists, like the Dutch, were assisted by their home government with provisions and equipment to found a settlement, this time on the lower reaches of the Mississippi River. They established the city of New Orleans in 1718 and called the area Louisiana, after Louis XIV. Early settlers there made their own fabrics and clothes but also bartered with the Indians for skins.

By 1700 Americans were dressing fashionably, and the distinctions between colonists of one nation and another were no longer very noticeable. Americans who were well-to-do followed the current fashions from Europe, and the main differences were between city dwellers and those from rural areas. Many of the latter still made their own clothes from homespun and woven fabrics, but the former could afford to import luxury fabrics and follow the fashion trends. Fashion dolls and costume plates now reached America, and it was possible to be au courant with the latest modes. Even during the years 1750–70, when luxurious styles prevailed in Europe, the Americans followed every frivolous idiosyncrasy.

In the first half of the 18th century, English colonists tended to follow English fashions, but the American Revolution altered this attitude. During the war there were severe restrictions on imported goods, and, when the war was over and independence had been won, most Americans did not return to buying their clothes from England; they went directly to the source of fashion—Paris.

(A.W.M./D.Y.)

THE OTTOMAN EMPIRE

From the early 12th century the Byzantine Empire had begun its slow decline in the face of the Turkish advance. In 1324 or 1326 the Ottoman Turks captured Bursa, on the opposite side of the Sea of Marmara from Constantinople, and this city became the first capital of the young empire. In 1453 Constantinople itself fell to the Turks. By then the costume of both cultures had been influencing one another for many years. From the Turks had come the wearing of the caftan and trousers; the Byzantines contributed beautiful silks, jeweled embroideries, and cloth of gold. The Turks adopted this richness of attire with such enthusiasm that, by the 16th century, sultans were trying to stem the tide of luxury in dress, as western Europe had long been attempting to do, by the passing of sumptuary laws forbidding the wearing of these materials and decoration except by the privileged few.

From the 15th century until the modernization of Turkey soon after 1918, the basic garments of the general population changed comparatively little. Although a Europeanization movement had begun about the middle of the 19th century, this was a slow process, affecting mainly the dress of the upper strata of society and that of the urban population. For many years such attire was a blend of styles from western Europe worn together with traditional Ottoman garments.

Traditional Ottoman dress

Traditional men's dress comprised a shirt, trousers, jacket, and boots. The trousers were of the very full, baggy type (similar to the Middle Eastern *chalvar*), fitting tightly only on the lower leg. A deep waist sash, the *kuşak*, bound the body over the junction between trouser and shirt. The jacket was a short one, worn open, and was decoratively embroidered. In cold weather a caftan would be worn on top of these garments. The only difference between the clothing worn by the average member of the population and those in a higher social class was that the garments of the latter would be made from richer, more decorative fabrics and that a long caftan would be worn on top. Sometimes more than one such coat was worn, with or without sleeves.

The traditional Turkish cap, the tarboosh, resembles an inverted flower pot and is made of cloth or felt. Similar to the fez, a term believed to have derived from the Moroccan town of that name, this cap was for centuries under the Ottoman Empire bound around the brow with a turban. The cap was made part of the national dress of the Turks during the 19th century and remained so until it was proscribed when Turkey became a republic in 1923. It is still worn by Muslims of both sexes in the Middle East.



Figure 26: Turkish woman wearing long *anteri* over patterned *chalvar* and backless slippers. Miniature from an album of single-figure studies, Turkey, 1618. In the British Museum.
Reproduced by courtesy of the trustees of the British Museum

The dress for women in the Ottoman Empire was very similar to that worn by Muslim women in the Middle East. It consisted of a knee-length, white, sleeved chemise (*gömlek*) and long drawers tied at the waist (*dislik*). The usual full trousers (*chalvar*) were accompanied, as in men's dress, by a decorative waist sash (*kuşak*). Over these garments a waistcoat (*yelek*) and long gown (*anteri*) were worn. Backless slippers were worn indoors. Outdoors the enveloping cloak (*tcharchaf*) and veil (*yashmak*) were obligatory, and decorative pattens (*kub-kobs*) kept the elegant slippers out of the mud of the streets.

The Ottoman Empire extended its control westward in Europe over the Balkans and as far as Hungary, where masculine dress was strongly influenced by Turkish styles. Until well into the 18th century men in these non-Muslim areas wore the dolman over the *mente* (both styles of caftan), together with trousers, boots, and a fur-trimmed hat known as the *kucsma*. Dress for women in these areas, however, followed the current styles of western Europe.

EUROPE AND AMERICA: 19TH AND 20TH CENTURIES

The 19th century. The influence of national features in dress had been declining since about 1675 and by 1800 had become negligible; from then on fashionable dress design was international. The character of the feminine wardrobe stemmed from Paris, the masculine from London. The English gentleman was established as the best-dressed, best-mannered in Europe, the lead being set by elegants such as Beau Brummell, whose clothes were copied by the prince regent himself (later King George IV). Brummell was so concerned with fit that he had his coat made by one tailor, his waistcoat by another, and his breeches by a third. The neckcloth was so elaborate and voluminous that Brummell's valet sometimes spent a whole morning getting it to sit properly.

Dominance of British menswear

It was during this period (*c.* 1811–20) that English modes for men became everywhere accepted as correct, even in Napoleonic France (the top hat, for example, became almost universal), and men's dress slowly became stereotyped, etiquette having laid down detailed regulations for the attire to be worn for different occasions, for different times of day, and by the various social classes. The tailcoat, waisted and padded on the chest, was de rigueur, accompanied by a waistcoat and now by close-fitting trousers called pantaloons, which were first buckled at the ankle and later, after 1820, strapped under the instep.

French dominance of what was chic for women was absolute during the 19th century. Parisian designs of garments and accessories were publicized throughout Europe and America by fashion plates and journals. At first originating from England and France, after 1850 they came from all European countries, and the Americans introduced some of the later world-famous journals—for example, *Vogue* and *Harper's Bazaar*.

Until about 1820 women's dress continued to reflect styles initiated as a result of the French Revolution. These fashions, which attempted to translate such ideals of the Revolution as democracy and liberation, were supposedly based upon the classical dress of ancient Greece. Ladies wore loose, draped, high-waisted gowns in white or pale colours in imitation of those depicted by white marble statuary. Corsets were abandoned; indeed, women wore a minimum of thin garments with little underwear—a totally unsuitable garb for the winters of northern Europe. To attempt to offset the chill, women adopted a three-quarter-length overdress made from a warmer, richer-coloured material and a variety of stoles, shawls, and pelisses. Hair was dressed classical fashion in a chignon bound with ribbons and decorated with plumes. The Romantic age of the 1830s brought back more colour, a tighter waistline at a more natural level, fuller, shorter skirts, leg-of-mutton sleeves, and complex high coiffures surmounted by large-brimmed hats or bonnets.

From the 1840s men's dress began to be more casual but lost most of its colour: black, shades of gray, and white were the norm. The formal black tailcoat was now reserved for evening attire. For daywear, tailcoats of various types were worn with a waistcoat and the new looser style of trousers over boots. Neckwear was plainer, consisting of a collar with neck scarf. The three-piece lounge suit, with a jacket instead of a tailcoat, was introduced in the 1850s for informal occasions. In the last two decades of the century a more countrified attire consisting of Norfolk jacket and knickerbockers became popular. (The name knickerbocker was taken from the nom de plume—

Introduction of the three-piece lounge suit

By courtesy of the Rijksmuseum, Amsterdam

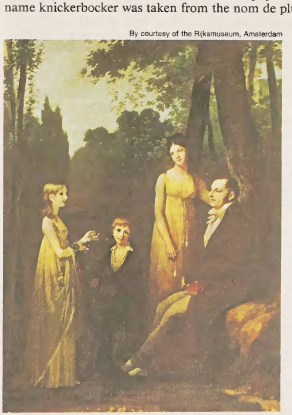


Figure 27: *Early 19th-century dress.*
Women wearing classical-style gowns with high waistlines and short sleeves; man in frock coat, waistcoat, and pantaloons; boy in skeleton suit and shirt. "The Schimmelpenninck Family," oil painting by Pierre Paul Prud'hon, 1801. In the Rijksmuseum, Amsterdam.





Figure 28: (Left) Women wearing light summer dresses with crinolines. "Women in the Garden," oil painting by Claude Monet, 1866–67. In the Louvre, Paris. (Right) Women wearing bustles and men in formal tailcoats. "Too Early," oil painting by James Tissot, 1873. In the Guildhall Art Gallery, London.

(Right) By courtesy of the Guildhall Art Gallery, London, photograph, A.C. Cooper Ltd.; (left) Giraudon—Art Resource/EB Inc.

Diedrich Knickerbocker—adopted by Washington Irving for the comic history of New York that he wrote in 1809. The Knickerbockers, a family of Dutch settlers in 17th-century New Amsterdam, were depicted in George Cruikshank's book illustrations wearing the full breeches of this time.)

The greater informality extended to hat design, with new styles being introduced that were gradually adopted for less informal use. The bowler, also known by such other names as the colloquial British "billycock" and, in America, the derby, was introduced about 1850 by the hatter William Bowler. The straw boater, originally meant to be worn on the river, became popular for all summer activities. The homburg felt hat, introduced in the 1870s and popularized by the Prince of Wales (later King Edward VII), stemmed from the German town of that name. Also popular at this time for sports and country wear in Britain was the deerstalker cap immortalized in Sir Arthur Conan Doyle's Sherlock Holmes stories.

Between about 1840 and 1870, long, bushy side-whiskers were fashionable. These whiskers, which left the chin clean-shaven, were called Piccadilly weepers in England; in America they were commonly referred to as burnsides or sideburns, after the U.S. Civil War general Ambrose Burnside, or as dundrearies, after a character in a contemporary play. Other popular beard styles included the imperial, a small goatee named for Napoleon III, and the side-whiskers and drooping mustache known as the Franz Joseph in honour of the head of the Austro-Hungarian Empire. After 1880 men tended to be clean-shaven or to wear a mustache only.

The second half of the 19th century was a time of prosperity in Europe. Despite wars and upheavals, the upper classes dressed fashionably and luxuriously. The styles of men and women acted as foil to one another—the men's dress sombre, dignified, and only slowly changing, that of the ladies changing ever faster in a kaleidoscope of modes. The technical advances and the capability for mass manufacturing brought about by the Industrial Revolution were making fashionable dress available to a rapidly expanding middle class. The invention of the sewing machine and the jacquard loom for weaving patterned textiles, the development of the ready-to-wear trade, the growth of new marketing techniques, and the establishment of department stores were revolutionizing the fashion scene.

Haute couture

In France haute couture had taken over control of the fashion-design world. The Englishman Charles Frederick Worth, who had emigrated to Paris in 1845, was the first of the great couturiers and one of the most influential. He introduced the practice of preparing a collection of designs, and he was the first to use live mannequins to display designs to buyers. Although only the rich could afford designer fashions, the styles gradually reached the ready-to-wear market (in a modified form that nonetheless prompted the introduction of new fashions for the upper classes), so that haute couture came to dictate women's fashions.

Women's dress from 1840 onward was, once more, dominated by a restrictive boned corset and framework underskirt. The fullness of the skirt was at first achieved by adding more layers of petticoats, leading to the crinoline petticoat of 1850. Named after the materials from which it was originally made (Latin: *crinis,* "[horse] hair"; *linum,* "thread"), this petticoat was, like its predecessors the farthingale and the hoop, a heavy underskirt reinforced by circular hoops of whalebone. By 1856 the weight of the crinoline and the petticoats became intolerable, and the cage crinoline was invented. This was a flexible steel framework joined by tapes and having no covering fabric. Sold at two shillings and sixpence, it was immensely popular and worn by most classes of society, at least for Sunday dress. It became the target for cartoonists, who took full advantage of all possible ludicrous situations; but this in no way lessened its popularity.

Gradually, in the 1860s, the shape of the crinoline changed, metamorphosing into that of the rear bustle, which was fashionable in the 1870s and '80s. Only in the 1890s did the skirt return to a relatively slender silhouette, but there was no letup in the constrictive corset, which was then at its most painful and harmful stage. In general, the styles of the late 19th century were feminine and elegant but not easy to wear. They restricted natural movement with their multiple layers, overdecoration, and sheer quantity of material.

Women's hair, always worn long during the century, was, from about 1840 to 1870, dressed in a severe style in which it was drawn back tightly from a centre parting into a bun at the back. Later styles were more feminine, dressed high on top and in a chignon or ringlets behind. The bonnet in many and varied guises was the chief head covering and was replaced by dainty hats only in the 1870s and '80s. Throughout the 19th century cosmetics were disapproved of; this was a strong reaction to the excessive use in the previous century.

Not all the women of the 1880s, however, wore these fashionable clothes. Followers of the Aesthetic movement in England wore looser garments—though the waists were still tight—with enormous sleeves supposed to resemble those worn by women in early Florentine paintings. The humorous journals of the period made great play with the contrast between fashionable and Aesthetic modes.

"Bloomers"

An earlier attempt to introduce a more comfortable, practical attire for women had been made by the American Elizabeth Smith Miller. The costume she designed was enthusiastically advocated by her friend Amelia Jenks Bloomer, a journalist and writer. In 1851 Bloomer traveled to London and Dublin to publicize this dress reform. The outfit, consisting of a jacket and knee-length skirt worn over Turkish-style trousers, was regarded as immodest and unfeminine. It was greeted with horror and disdain, and the idea quickly died. What has survived is the name "bloomers," which originally referred to Miller's full trousers but was later applied to long knickers worn as underwear in the early 20th century. Miller's garment was also the inspiration for "rationals" (sometimes also known as bloomers), the knickerbockers worn by women for cycling and sport in the 1890s.

Children's clothes were less sensible and comfortable than they had been 50 to 60 years before. What had started in the 1820s as rational dress for boys had been formalized into the rigid discomfort of the Eton suit with its stiff white collar. Fortunate boys were dressed in sailor suits, and unfortunate ones as "Little Lord Fauntleroy," in velvet suits with lace collars and cuffs and with the hair dressed long in curls. Little girls were dressed in elaborate and easily soiled garments with much lace. Their skirts were shorter than those of adult women, but the waists were nearly as tight. (Ja.L./D.Y.)

The early 20th century. There were no fundamental changes in dress during the first decade of the 20th century. Men continued to wear a black frock coat with gray striped trousers for formal day wear and a black tailcoat and trousers with a white waistcoat for evening wear if ladies were present, although in America the tuxedo, or dinner, jacket was beginning to provide a more comfortable alternative. (The term derives from the fact that the style was introduced in the millionaire district of Tuxedo Park in the state of New York for wear at small dinner parties.) Three-piece lounge suits were worn for less formal day functions, and for country and sportswear the Norfolk jacket and knickerbockers remained popular.

For ladies the ideal figure was an unusual one, comprising a full, forward-thrusting bosom, tiny waist, and generous, backward-slanting hips. This unnatural S-bend posture was achieved mainly by a boned corset that was long and rigid in front and shorter at the rear. The costume was extremely feminine, overdecorated with flounces and lace, frills and embroidery. It was totally impractical but ornamental and attractive. Picture hats were set upon pompadour coiffures, affixed with vicious-looking hatpins. The high neckline was boned at the sides up to the ears, and the skirt hem trailed on the ground, its beautiful fabric protected from the dirt by ruffled petticoat "sweepers."

From 1910 important changes began to take place in feminine attire. The French couturier Paul Poiret led the field in designing an exotic range of glamorous creations made from superb fabrics in brilliant colours. Poiret, who was designing for the "new woman," freed women from the multiplicity of petticoats and from the excruciating corset. His gowns still reached the ground, however, and the skirts were restrictive, making it difficult to walk. His hobble skirt, in which the material was very narrow at the ankle, was particularly aptly named; in some cases a deep band encircled the skirt at ankle level, rendering it difficult to put one foot in front of the other. Poiret's other designs included the lampshade, or hoop tunic, skirt. He also took many of his more exotic designs from Oriental prototypes; these included turbans, Eastern-style trousers, and harem skirts.

Influence of social changes on dress

Women were beginning to question their status in a man's world. Some became suffragettes, some went to work outside the home. A more practical form of dress became popular, with the blouse and skirt replacing the ruffled tea gown. During the war years of 1914–18 these changes accelerated. A minority of women were in uniform, but far more worked in factories, in offices, as postal carriers and in other jobs previously performed by men. To meet their needs, the picture hat was replaced by a small neat design, and the skirt hemline rose to eight inches above the ground, revealing the ankles for the first time.



Figure 29: *Turn-of-the-century fashions.* (Top) Women wearing rationals, men's stiff collars, and straw boaters for cycling. "The Cycle Hut in the Bois de Boulogne," oil painting by Jean Beraud, *c.* 1901–10. In the Musée de l'Ile de France, Château de Sceaux. (Bottom) Women dressed in cascades of lace, with corsets that threw the hips back and the bosom forward in the S-shaped look; men in frock coats, white waistcoats, stiff white collars, and silk top hats or straw boaters. "Jardin de Paris. The Night Beauties," oil painting by Jean Beraud, 1905. In the Musée Carnavalet, Paris.

(Bottom) By courtesy of the Musee Carnavalet, Paris; photograph, (top) Cliche Flammarion/Musee Chateau de Sceaux, France

The postwar 1920s brought a complete change to the fashion scene. Men's dress moved more slowly than women's, but even here far more variety appeared in colour schemes and fabrics, and the trend toward informality was accelerated. The tailcoat was reserved for weddings and dances, the lounge suit became the accepted city wear, and sports jackets and gray flannels were popular for casual attire. After 1925 trousers commonly featured turnups (cuffs in America), and the legs became increasingly wider; the popular "Oxford bags" measured 20 inches at the hem. Knickerbockers had become fuller and longer, overhanging the kneeband by four inches, and were thus known as plus fours, which remained fashionable until at least 1939. Knitted pullovers (often homemade) in coloured (fair isle) patterns replaced the waistcoat for informal occasions. Technical advances had improved water-repellent fabrics, and most men had a raincoat. A favourite style



Figure 30: "Afternoon Dress with Hoop Tunic," by Bernard Boutet de Monvel, 1914. In the collection of François Boucher, Paris.

In the collection of F. Boucher, Paris, permission A.D.A.G.P. 1972 by French Reproduction Rights, Inc., photograph, Flammarion

was the trench coat, a classic design based upon the coats worn by World War I officers in the trenches. Men were mostly clean-shaven, and their hair was short. A peaked cap accompanied leisure wear, and a trilby felt hat the lounge suit. (The latter was named after George du Maurier's novel; the American term was fedora, named for the heroine of a play.)

For women in the 1920s, freedom in dress reflected the new freedoms opening up for them to take up careers, to study at college, to enter professions. Only a small percentage of women took up such opportunities, far fewer than today, but the revolutionary changes nonetheless affected the type of clothes worn by most women in the Western world. The skirt hemline rose steadily to become, at its shortest in the years 1925–27, knee-length. With the short skirts, flesh-coloured stockings were introduced, made from expensive silk or more practical lisle or wool (other colours were also worn). Corsets, layers of petticoats, and overdecoration all disappeared to be replaced by a boyish figure style in which the waist, breasts, and hips were all understressed. Probably the most revolutionary change was in the coiffure. Hair was first cut shorter during the war years to make it easier to care for, and by the 1920s the shingle and the more severe Eton crop were being adopted by the ultrafashionable. Marcel waving, introduced in the late 19th century, and the later "perm" also contributed to manageability and became popular at this time. Permanent waving abolished forever the nightly chore of setting curl papers and pins. The new, sleek hairstyles were accompanied (and virtually concealed) by the generally unflattering cloche hat, which closely covered the head down to the eyes: only a beautiful woman could wear one with chic.

Women's hairstyles

Femininity returned in the 1930s. The ideal figure was still slim, but the low-waisted, tubular look was out of style. The uplift bra enhanced the breasts, and the waistline returned to its natural level. The skirt lengthened again until it reached about eight inches above the ground for the daytime and ground length for the evening. For evening styles, the backless dress and halter neckline became fashionable. The bias cut of material, a mode introduced in the 1920s by the French couturiere Madeleine Vionnet, was widely adopted in the 1930s and was very effective with the longer skirts, creating a figure-hugging style which then flared out at the hemline.

By this time there was available a great variety of specialized clothing for different occasions, including for sport and leisure or resort activities, such as swimming, walking, and cycling, as well as for parties and dances. The cosmetics industry, led by the United States, had also expanded and became big business in the 1930s; most women routinely carried face powder, lipstick, eye shadow, and tweezers in their handbags for running repairs.

Influence of World War II on dress

The high proportion of men and women in uniform in the years 1939–45 strongly influenced the civilian dress style. For women, garments had square padded shoulder lines, and skirts were a practical knee length. Trousers were widely worn by both civilian and military women. After World War II, trousers and trouser suits remained popular, especially between 1945 and 1970. In Europe, the war years meant austerity and clothing coupons; fashion did not have a high priority. Shortages of materials both during and immediately after the war led to the introduction of "utility" styles, especially in Britain, where gov-

Courtesy *Vogue*. Copyright © (left) 1928 (renewed 1956), (centre) 1931 (renewed 1959), (right) 1947 by the Conde Nast Publications Inc.



Figure 31: *Women's fashions from the 1920s to the 1940s.*
(Left) Tubular, long-waisted, short-skirted dress. *Vogue*, 1928. (Centre) Bias-cut evening dresses with flared hemlines, drawing by Bolin, *Vogue*, 1931. (Right) Christian Dior's "New Look," with narrow shoulders and long skirt. *Vogue*, 1947.

ernment rulings insisted on the removal of all superfluous trimmings, including pockets and pleats, and restricted the fullness of garments in order to economize on the amount of fabric used.

Post-World War II. It is extremely difficult to sum up the changes in dress that have taken place since 1945. Fashions have changed at a far greater pace than ever before, a pace that is still accelerating. The rules of etiquette governing what type of dress should be worn by whom and when have virtually disappeared. It has become the accepted dictum to "do your own thing," to choose clothes, whether for day or evening, formal or holiday wear, according to personal inclination. Wide-scale advertising, especially on television, and the modern marketing system have brought fashion within the reach of all, both in cost and availability. Leading manufacturers and department stores purchase original designs from fashion houses and then manufacture ready-to-wear versions in quantity at various price levels to suit the entire population.

One of the most influential factors in the development of modern fashions has been the technological advance in the production of synthetic textile fibres. Permanent pleating, fast dyes, crease resistance, preshrinking, and other easy-care characteristics of synthetics have made it possible to manufacture clothing more quickly and less expensively. Although traditional natural fabrics remain popular, they have been almost completely replaced by synthetics in the manufacture of some garments. Women's stockings made of nylon, for example, first went on sale about 1940 and, after World War II, soon supplanted all other types. Similarly, the underwear industry was revolutionized when latex thread was employed, along with the zipper, to fabricate comfortable two-way stretch suspender belts, effectively banishing the hated corset.

The keynote of the changes in men's dress has been casualness. The tailored jacket and vest have been steadily ousted and often replaced by knitted pullovers and cardigans. Central heating of homes and transport by car have virtually done away with overcoats, heavy tweed suits, and hats; well-cut shirts and trousers are normal office and car wear. In line and cut, fashion styles have changed more quickly than ever before, with the narrow cuffless trousers, trousers with the waistband at hip level, and the bell-bottomed flared trousers all popular at various times. For elegant evening wear a coloured velvet jacket with cummerbund was long favoured, although many men accepted little distinction between day and evening attire.

Soon after the war the French designer Christian Dior raised feminine morale after years of drabness by introducing his 1947 "Corolle" collection, soon to be dubbed the "New Look" by the American press. Here was a return to femininity: a long, full skirt with a bouffant ruffled petticoat beneath, a slender waist, and sloping shoulders. The look did not last long. Although women liked it, it was not sufficiently practical for the new world of women out at work.



Figure 32: *Men's dress, early 1970s.*
Men wearing "mod" suits, wide-legged trousers, broad ties, and a trench coat with casual turtleneck sweater, 1971.
Reprinted, courtesy of the *Chicago Tribune*

More popular, so much so that the fashion was slow to die and has since been revived, was the miniskirt, introduced in 1965 by Mary Quant in London and André Courrèges in Paris. Starting at mid-thigh length, the hemline crept upward to become a micro-skirt, a style that had only been made feasible by the introduction of nylon

The miniskirt

(Left) Pictorial Parade, (right) UPI/Bettman





Figure 33: *Women's dress of the 1960s and '70s.*
(Left) Miniskirt outfit with clump-heeled shoes, 1969. (Right) Polyester and wool pantsuits, 1970.

tights (panty hose in the United States). Other lengths appeared—the midi and the maxi—but neither was as popular as the mini.

A feature of fashion since 1945 had been the emphasis on clothes for the young, something never before experienced. Throughout history children and young people wore basically the same type of clothes as their parents. After 1945 a generation was growing up to enter a world of easy employment opportunities and good wages. The marketers of clothes took full advantage of this and aimed their designs toward the young; a complete teenage wardrobe evolved, comprising garments almost unwearable by older people. Clothes were extremely tight-fitting and casual. Blue jeans became and, indeed, continued to be a uniform for the young. Young men and women began to ape each other's styles, and unisex clothes were born. In the 1960s London's Carnaby Street became an important centre for antiestablishment "mod" fashions. Since then, styles have moved quickly and have been full of contradictions. Ethnic, romantic, nostalgic, erotic, punk, and conservative effects, among others, have all had their adherents.

(D.Y.)

The history of Eastern dress

Western-style clothes, which many people find convenient to wear during business hours, are now a common sight in many large cities of eastern and southern Asia. This is particularly so in Japan, a country which, since 1945, has been especially influenced by the American way of life and has built a reputation as an international fashion centre. However, even here, as in much of Asia, it is not uncommon for a reversion to traditional dress to take place in the home.

Over the centuries, notably in Korea and Japan, these traditional styles of dress have reflected marked Chinese influence, though both countries developed characteristic styles of their own. In like manner, modes of dress in the Indian subcontinent have been a source of inspiration to some of the countries of Southeast Asia and of the East Indian archipelago.

CHINA

More than 2,000 years before the beginning of the Christian era, the Chinese discovered the marvelous properties of silk and shortly thereafter invented looms equipped with devices that enabled them to weave patterned silks rapidly enough to satisfy the demand for them by luxury-loving Chinese society. Thus, centuries before Chinese silks began to be shipped westward and still more centuries before the West learned the secret of sericulture, the people of China had already established ultrarefined standards of elegance in matters of dress.

The earliest period of Chinese history for which reliable visual evidence of clothing styles is obtainable is the Han dynasty (206 BC–AD 220). Han bas-reliefs and scenes painted in colour on tiles and lacquers show men and women dressed in wide-sleeved kimono-style garments which, girdled at the waist, fall in voluminous folds around their feet. The graceful dignity of this *p'ao*-style robe, which continued to be worn in China until the end of the Ming dynasty in 1644, is clearly revealed in Chinese figural paintings attributable to the interval between the 8th and the 17th century (Figure 34). Other traditional garments include the tunic or jacket, worn by both sexes over loosely cut trousers. For colder weather, clothing was padded with cotton or silk or lined with fur.

The *p'ao*

Chinese records indicate that at least as early as the T'ang dynasty (618–907) certain designs, colours, and accessories were used to distinguish the ranks of imperial, noble, and official families; but the earliest visual evidence of these emblematic distinctions in dress is to be found in Ming portraits. In some of these, emperors are portrayed in voluminous dark-coloured *p'ao* on which the 12 imperial symbols, which from time immemorial had been designated as imperial insignia, are displayed. Other Ming portraits show officials clothed in red *p'ao* that have large bird or animal squares (called "mandarin squares," or *p'u-fang*) on the breast, specific bird and animal emblems



Figure 34: The legendary Emperor Yao wearing a *p'ao*-style robe, a wide-sleeved kimono-style garment, which, girdled at the waist, falls in voluminous folds at the feet. Hanging silk scroll by Ma Lin, Southern Sung dynasty (1127–1279). In the National Palace Museum, Taipei, Taiwan.

By courtesy of the National Palace Museum, Taipei, Taiwan, Republic of China

to designate each of the nine ranks of civil and military officials having been adopted by the Ming in 1391.

When the Manchus overthrew the Ming in 1644 and established the Ch'ing dynasty, it was decreed that new styles of dress should replace the voluminous *p'ao* costume. The most formal of the robes introduced by the Manchus was the *ch'ao-fu*, designed to be worn only at great state sacrifices and at the most important court functions. Men's *ch'ao-fu* (Figure 35) had a kimono-style upper body, with long, close-fitting sleeves that terminated in the "horsehoof" cuff introduced by the Manchus, and a closely fitted neckband over which was worn a detached collar distinguished by winglike tips that extended over the shoulders. Below, attached to a set-in waistband, was a full, pleated or gathered skirt. Precisely stipulated colours and pattern arrangements of five-clawed dragons and clouds, waves, and mountains were specified for the *ch'ao-fu* of emperors, princes, nobles, and officials; the bright yellow of the emperor's robe and the 12 imperial symbols emblazoned on it clearly established his lofty rank. All other ranks wore "stone blue" *ch'ao-fu* decorated in accordance with prescribed rules about the number, type, and arrangement of dragon motifs.

Only women of very high rank were permitted to wear *ch'ao-fu*. Women's robes were less commodious than the men's and were cut in long, straight lines with no break at the waist. The narrow sleeves with horsehoof cuffs of these *ch'ao-fu* robes and the arrangement of their dragon, cloud, mountain, and wave patterns were essentially the same as those of the so-called dragon robes discussed below. They were clearly differentiated from the dragon robes, however, by their capelike collars and by flaring set-on epaulets which, gradually narrowed, were carried down under the arms. Stolelike vests, always worn over women's *ch'ao-fu*, were also a distinguishing feature of this costume.



Figure 35: *Ch'ao-fu* worn by Ho-shen, minister to the Ch'ien-lung emperor (reigned 1735–96). Painted silk wall hanging. In the Metropolitan Museum of Art, New York City.

By courtesy of the Metropolitan Museum of Art, New York, Rogers Fund, 1942

"Dragon robes"

Chi-fu, or "dragon robes" (*lung-p'ao*) as they were usually called, were designed for regular court wear by men and women of imperial, noble, and official rank. The *chi-fu* was a straight, kimono-sleeved robe with a closely fitted neckband that continued across the breast and down to the underarm closing on the right side, the long tubular sleeves terminating in horsehoof cuffs. The skirt of the *chi-fu* cleared the ground to permit easy walking and in men's garments was slit front and back as well as at the sides to facilitate riding; the extra slits were the only feature that distinguished the *chi-fu* of men below the rank of emperor from those of their wives. All *chi-fu* were elaborately patterned with specified arrangements of dragons, clouds, mountains, and waves, to which were added auspicious or Buddhist or Taoist motifs. Distinctions in rank were indicated by the colours of the robes and by slight variations in the basic patterns; however, because of the large number of personages who wore *chi-fu,* these distinctions were not always easily recognizable. Emperors' *chi-fu,* either yellow or blue, were always distinguished by the 12 imperial symbols.

The informal Manchu *ch'ang-fu,* a plain long robe, was worn by all classes from the emperor down, though Chinese women also continued to wear their Ming-style costumes, which consisted of a three-quarter-length jacket and pleated skirt. Men's *ch'ang-fu,* cut in the style of the *chi-fu,* usually were made of monochrome patterned damask or gauze; women's *ch'ang-fu* had wide, loose sleeves finished off with especially designed sleevebands decorated with gay woven or embroidered patterns.

The declining Ch'ing dynasty was finally swept aside in 1912, and Western influences exerted pressure on China to begin to emulate the world outside its boundaries. Under the new republic the traditional Chinese culture began to give way to modern ideas. Gradually this was reflected in dress. By the 1920s women, in particular, adopted a compromise attire. This was the *ch'i-p'ao,* better known in the West by its Cantonese name, cheongsam. The *ch'i-p'ao* had developed from the *ch'ang-fu,* and by 1930 the majority of women were wearing it. A close-fitting dress made from one piece of material, the *ch'i-p'ao* was fastened up the right front side. It had a high mandarin collar, and its skirt was slit up the sides to the knee. It was made of traditional Chinese fabrics, padded in winter for warmth. At first it was a long dress, but the hemline gradually rose to come into line with Western dress.

In mainland China the communist revolution of 1949 brought strict directives on dress. Styles were to be the same for everyone, whether man or woman, intellectual or manual labourer. This drab uniform was a blend of peasant and military design. It consisted of a military-style high-collared jacket and long trousers. Men's hair was short and covered by a peaked cap. Women's hair was longer but uncurled. Shoes had flat heels. No cosmetics or jewelry was permitted. Traditional Chinese cotton was used to make the garments; colour designated the type of worker. After about 1960 a slow Westernization set in, permitting a variation in colour and fabric. Dresses were introduced for women.

JAPAN

Earliest representations of Japanese dress styles

The earliest representations of dress styles in Japan are to be found in 3rd- to 5th-century-AD clay grave figures (*haniwa*), a few of which show men and women wearing meticulously detailed two-piece costumes consisting of crossed-front jackets that flare out over the hips, the men's worn over full trousers, which, banded above the knees, hang straight and loose beneath; women's jackets were worn over pleated skirts (Figure 36).

Two-piece costumes appear to have been worn regularly during the 7th and 8th centuries, the jackets of this period being called *kinu,* the men's trousers *hakama,* and the women's skirts *mo.* It is known, however, that during the Nara period (710–784) Japanese court circles adopted Chinese court dress, the most characteristic feature of which was the long kimono-style *p'ao* garment; thus, it must be supposed that the *kinu, hakama,* and *mo* were the accoutrements of middle- and lower-class society, though these garments may also have been adapted for wear under the *p'ao.* It is clear that emblematic colours and patterns as well as the *p'ao* style were borrowed from China because modern court dress in Japan, which has been little changed since the 12th century, has many purely Chinese characteristics.

Court costumes

The most important court costumes of Japan are the *sokutai* of the emperor and the *jūni-hitoe* of the empress, which are worn only at coronations and at very important ceremonial functions. (Similar costumes are worn by the crown prince, by princes and princesses of the blood, by

By courtesy of the National Museum, Tokyo



Figure 36: Japanese grave figure, wearing meticulously detailed two-piece costume of flared, crossed-front jacket and pleated skirt. Clay, 3rd to 5th century. In the National Museum, Tokyo.

high officials, and by ladies-in-waiting.) The voluminous outer robe (*ho*) of the emperor's *sokutai* is cut in the style of the Chinese *p'ao* but is given a distinctively Japanese look by being tucked up at the waist so that the skirt ends midway between the knees and the floor. This *ho* robe is yellow (the colour worn only by emperors and their families in China), and it is patterned with *hō-ō* birds and *kilin* (Japanized versions of the mythical Chinese *feng-huang* and *ch'i-lin*). The outer and most important of three kimonos worn under the *ho* is the *shitagasane,* which has an elongated back panel that forms a 12-foot (4-metre) train. The *shitagasane* is made of white damask, as are the baggy white trousers (*ue-no-hakama*) that are a characteristic feature of the *sokutai* costume. Both of these garments and a cap-shaped headdress (*kammuri*) of black lacquered silk, with an upright pennon, decorated with the imperial chrysanthemum crest, are purely Japanese in style, but the ivory tablet (*shaku*) carried by the emperor when wearing the *sokutai* was undoubtedly inspired by tablets of jade that Chinese emperors carried as symbols of their imperial power.

The outermost garment of the empress' *jūni-hitoe* costume is a wide-sleeved jacket (*karaginu*) that reaches only to the waist and has a pattern of *hō-ō* bird medallions brocaded in colours of the empress' choice. Attached to the waist at the back of the *karaginu* is a long, pleated train (*mo*) of sheer, white silk decorated with a painted design. The outer kimono (*uwagi*) is very large to accommodate the many layers of kimono worn under it, the abnormally long skirt swirling out fanwise around the wearer's feet. This, too, is made of rich brocade, its design and colours being a matter of personal taste. Under the *uwagi* is a plain purple kimono, and under that a robe known as the *itsutsu-ginu,* which has multiple bands of coloured silks (usually five) attached at the edges of the sleeves, at the neckline, and at the hem, giving the appearance of several robes worn one over another. No special interest attaches to the *hitoe* kimono worn under the *itsutsu-ginu* or to the *kosode* worn next to the body, but the divided skirt (*naga-bakama*) that completes the costume is an extremely picturesque garment. Made of stiff, red cloth and fastened high up under the breasts, the *naga-bakama* covers the feet in front and is carried out in a train in back. Worn with the *jūni-hitoe* is an elaborate coiffure known as *suberakashi,* and affixed directly over the forehead are special hair ornaments consisting of a lacquered, gold-sprinkled comb surmounted by a gold lacquered chrysanthemum crest.

Other types of dress formalized in the 12th century were the *noshi* (courtiers' everyday costumes) and the *kariginu,*



Figure 38: Japanese woman wearing an *uchikake*-type kimono. "The Courtesan Itsutomi Holding a Plectrum," print by Chobunsai Yeishi, c. 1794.
By courtesy of the Victoria and Albert Museum, London; photograph, A.C. Cooper Ltd.

worn for hunting. Both of these garments were voluminous hip-length jackets worn with baggy trousers tied at the ankles. At this time also it became necessary to devise special costumes for the newly formed samurai caste. The *hitatare,* the formal court robe of samurai, and the *suo,* a crested linen robe designed for everyday wear, were characterized by V-shaped necklines accentuated by inner-robe neckbands of white. Several centuries later the samurai adopted the *kamishimo,* a striking jumperlike garment, with extended shoulders and pleated skirt-trousers, which was worn over the *hitatare.* This costume probably inspired a later fashion of wearing skirt-trousers (*hakama*) over a full-length black kimono, which, together with the

Other types of 12th-century formalized dress

By courtesy of Kumata Shrine, Osaka, Japan



Figure 37: Two Japanese men and a woman (foreground) wearing the *kamishimo,* a jumperlike garment with extended shoulders; woman (far right) and men (background) in the *yukata,* a cotton kimono; wealthy merchant (seated right) wearing a short, black *haori* coat and *hakama,* men's skirt-trousers. Detail of "Shuin-Boeki-sen," a folding screen, late 17th century. In Ōsaka Castle, Japan.

short black *haori* coat, was until fairly recently the approved formal attire for Japanese men (Figure 37).

The basic kimono style adopted by Japanese women during the Nara period has remained amazingly close to that of the *p'ao* robes worn by the women of T'ang China. The practice of wearing a short-sleeved kimono (*kosode*) as an outer garment and belting it in with a narrow sash (*obi*) originated during the Muromachi period (Ashikaga shogunate; 1338–1573), when samurai women began to wear a voluminous outer kimono (*uchikake*) as a kind of mantle (Figure 38). Eventually, the *kosode* came to be worn only by married women, the long-sleeved *furisode* being reserved for young unmarried girls. The wide *obi*, tied in a variety of ways and fastened with an often intricately carved toggle (*netsuke*), was adopted in the early 18th century, and it was at this time also that women first began to wear the short *haori* coat, which has come to be an important feature of Japanese women's dress.

The *yukata*

The *yukata*, which is worn by both men and women, is a cotton kimono with stencil-dyed patterns (usually done in shades of indigo) that was originally designed for wear in the home after a bath. Because it has become accepted practice to wear *yukata* on the street on warm summer evenings, the cottons designed for them have become increasingly handsome.

Traditional Japanese footwear includes sandals, slippers, and wooden clogs (*geta*) worn with the *tabi*, a sock with a separate section for the big toe. (Pa.S./D.Y.)



Figure 39: Korean women wearing short *chŏgori*, a jacket, with raised up skirts. Inked and coloured leaf from an album of 30 leaves by Sin Yun-bok (1758–?). In the collection of Chun Sung-woo, Korea.
By courtesy of Chun Sung-woo, Korea

KOREA

Some of the basic elements of modern traditional dress in Korea, the *chŏgori* (jacket), *paji* (trousers), and *turumagi* (overcoat), were probably worn at a very early date, but the characteristic two-piece costume of today did not begin to evolve until the period of the Three Kingdoms (*c.* 57 BC–AD 668). During the early part of this period both men and women wore tight, waist-length jackets and short, tight trousers; and it is believed that the Koreans' traditional fondness for white clothing dates from this period.

Korean records state that special costumes for court wear modeled after those of T'ang China were adopted during the reign of Kim Ch'unch'u in the 7th century; but Chinese influence on Korean dress at this period is verifiable only in changes that occurred in the everyday costumes of the nobility. Noblewomen formerly had worn tight trousers and jackets (which continued to be worn by the poorer classes); now they began to appear in wide-sleeved, hip-length jackets, belted at the waist, and in full-length skirt-trousers. The corresponding dress for noblemen was a narrower, tunic-style jacket, cuffed at the wrists, belted, and worn with roomy trousers bound in at the ankles. The most striking evidence of Chinese influence at this time is to be seen in the style of the *turumagi* overcoat worn by noblemen, pictured in fresco paintings as a voluminous full-length garment made almost exactly like the *p'ao* robe of T'ang China. One-piece robes were never worn in Korea until the late 13th century, when the court was forced to adopt Mongol dress; after Mongol domination ended in 1364, Koreans wore the one-piece robe only at wedding ceremonies.

Everyday costumes of the nobility

In the 15th century, Korean women began to wear pleated skirts (*ch'ima*) and longer *chŏgori* (jackets), a style that was undoubtedly introduced from China. Noblewomen wore full-length *ch'ima* to indicate their social standing and began gradually to shorten the *chŏgori* until eventually it attained its present length, just covering the breast. This style made it necessary to reduce the fullness of the skirt somewhat in order to make it possible to extend it almost up to the armpits, which remains the fashion (Figure 39).

The adoption of Chinese-style mandarin squares as emblems of rank for civil and military officials (who wore them on their *turumagi*) appears to have been the only notable example of Chinese influence on men's dress at this period. Otherwise, few changes were made until 1894, when class distinctions were relaxed by government decree. It was at this time that the *turumagi* was shortened and narrowed to its present form.

The most picturesque costume of modern Korea is that of men of leisure, *yangban*, who are past 60 years of age. The *yangban* wear white almost exclusively, their costumes consisting of full trousers tied at the ankles with ribbons, over which is worn a short *chŏgori* and a fitted vest and, over all, a loose *turumagi*, which falls just below the knees and is tied at the breast. The patriarchal appearance of the *yangban* (who is usually bearded) is accentuated by a black horsehair hat, its flat brim and high crown giving him somewhat the appearance of an American colonial Pilgrim Father. Younger men wear a similar costume (though not the hat) in gray, light blue, or light brown.

Women's costumes feature a bolero-style white *chŏgori*, finished off at the neck by a figured band or ribbon that ties from left to right, and high-waisted *ch'ima*, which, in formal costumes, is a full, billowing garment made of beautifully patterned silk.

SOUTH ASIA

The Hindu population of South Asia comprises about 2,000 castes whose members wear clothes and ornaments that clearly indicate their caste. The subject of dress, therefore, cannot be dealt with satisfactorily in a few paragraphs. Some of the principal features of upper-class Hindu and Muslim dress and the history of their development can, however, be sketched briefly.

The ancient origin of two of the most characteristic garments of modern India, the dhoti worn by men and the sari worn by women, is verifiable in sculptured reliefs as far back as the 2nd century BC. Both men and women are pictured wearing a long piece of cloth wrapped around the hips and drawn between the legs in such a fashion that it forms a series of folds down the front. The upper bodies of both men and women were unclothed, though women wore a narrow cloth girdle around the waist. Men are pictured wearing large turbans, women with head scarves that fall to the hips. Women also wore a great amount of jewelry—bracelets, anklets, and girdles—but men's ornaments consisted solely of bracelets.

The dhoti and sari

No major change in costume appears to have been made until the 12th century, when the Muslims conquered northern and central India. In this part of the subcontinent, radical new dress styles were adopted to conform with Muslim practice, which required that the body be covered as completely as possible. Men's costumes thereafter consisted of the *jāmah*, a long-sleeved coat that reached to the knees or below and was belted in with a sash, and wide trousers known as *isar* (Figure 40, right). These garments and the *farjī*, a long, gownlike coat with short sleeves, which was worn by priests, scholars, and high officials, were made of cotton or wool, silk being forbidden to men by the Qur'ān. Somewhat modified, these

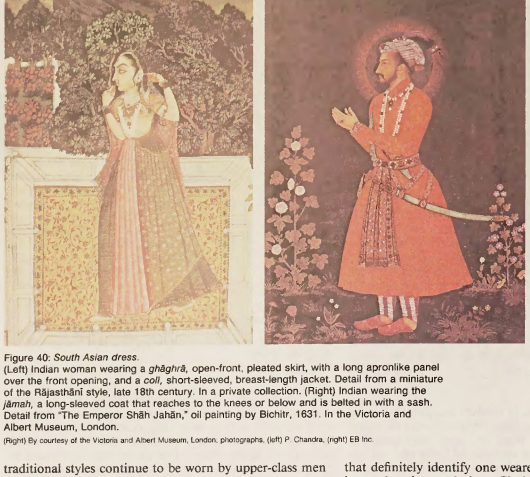


Figure 40: *South Asian dress.*
(Left) Indian woman wearing a *ghāghrā,* open-front, pleated skirt, with a long apronlike panel over the front opening, and a *coli,* short-sleeved, breast-length jacket. Detail from a miniature of the Rājasthānī style, late 18th century. In a private collection. (Right) Indian wearing the *jāmah,* a long-sleeved coat that reaches to the knees or below and is belted in with a sash. Detail from "The Emperor Shāh Jahān," oil painting by Bichitr, 1631. In the Victoria and Albert Museum, London.
(Right) By courtesy of the Victoria and Albert Museum, London; photographs, (left) P. Chandra, (right) EB Inc.

traditional styles continue to be worn by upper-class men of Pakistan and Bangladesh.

Muslim women's garments

Women's garments, dictated by the Muslim conquerors, consisted of wide-topped trousers snugly fitted around the calves of the legs, a long shirtlike garment, and a short, fitted outer jacket. Silk was not forbidden to women; and highborn women, forced to spend their lives in seclusion, devoted much time and money to their costumes. The Mughal emperor Akbar's Rājput wives, inspired by the profusion of luxurious fabrics available in India, designed a graceful new style of dress, which Muslim women adopted forthwith. This costume consisted of an open-front pleated skirt, or *ghāghrā,* worn with a long apronlike panel over the front opening, and a short-sleeved, breast-length blouse called a *coli* (Figure 40, left). The *ghāghrā* and *coli* continue to be basic elements of Muslim women's dress, the loose front panel replaced by the traditional sari, which is worn as an overgarment, one end draped around the hips, the other carried up over the shoulder or head.

Dress in southern India was little affected by Muslim rule in the north. The dhoti continued to be worn by most Hindu men (it is forbidden to some castes), and the sari by women. Some additions to these traditional costumes have been adopted. On formal and semiformal occasions many Hindu men wear a long, full-skirted, white cotton coat, which reaches to the knees and buttons down the front from top to bottom, over jodhpur-style white trousers; and most Hindu women wear a short *coli*-style blouse under a sari or with a long skirt under a loose waist-length bodice. (Pa.S.)

The nature and purposes of dress

Perhaps the most obvious function of dress is to provide warmth and protection. Many scholars believe, however, that the first crude garments and ornaments worn by humans were designed not for utilitarian but for religious or ritual purposes. Other basic functions of dress include identifying the wearer (by providing information about sex, age, occupation, or other characteristic) and making the wearer appear more attractive. Although it is clear why such uses of dress developed and remain significant, it can often be difficult to determine how they are achieved. Some garments thought of as beautiful offer no protection whatsoever and may in fact even injure the wearer. Items that definitely identify one wearer can lose their meaning in another time and place. Clothes that are deemed handsome in one period are declared downright ugly in the next, and even uniforms—the simplest and most easily identified costume—are subject to change. What are the reasons for such changes? Why do people replace useful, attractive garments before they are worn out? In short, why does fashion, as opposed to mere dress, exist?

Influence of social and economic standing on fashion

There are no simple answers to such questions, of course, and any one reason is influenced by a multitude of others, but certainly one of the most prevalent theories is that fashion serves as a reflection of social and economic standing. Thus, in relatively static societies with limited movement between classes, as in many parts of Asia until modern times or in Europe before the Middle Ages (or later in some areas), styles generally did not undergo major or rapid change. In contrast, when lower classes have the ability to copy upper classes, the upper classes quickly instigate fashion changes that demonstrate their authority and high position. During the 20th century, for example, improved communication and manufacturing technology enabled new styles to trickle down from the elite to the masses at ever faster speeds, with the result that more styles were introduced than at any other time.

Furthermore, the idea that fashion is a reflection of wealth and prestige can be used to explain the popularity of many styles throughout costume history. For example, courts have been a major source of fashion in the West, and clothes that are difficult to obtain and expensive to maintain have frequently been at the forefront of fashion. Ruffs, for example, required servants to reset them with hot irons and starch every day and so were not generally worn by ordinary folk. As such garments become easier to buy and care for, they lose their exclusivity and hence much of their appeal. For the same reason, when fabrics or materials are rare or costly, styles that require them in excessive, extravagant amounts become particularly fashionable—as can be seen in the 16th-century vogue for slashing outer garments to reveal a second layer of luxurious fabric underneath.

Similarly, impractical fashions that clearly demonstrate the wearer does not need to work, and indeed would find it difficult to do so dressed in such a manner, have often been considered beautiful. Examples include the Chinese practice of binding aristocratic women's feet, making it

impossible for the women to walk far, and the recurrent popularity in Europe of styles that limited a woman's ability to maneuver or move by confining her into frequently injurious corsets and weighting her down with excessive layers of petticoats and skirts. Women have traditionally been the targets of the most extreme forms of impractical fashion because they have frequently been viewed as little more than a frivolous ornament for a man's arm or household. The fact that a woman is dressed in such a manner proves not only that she does not work but also that her husband or father can afford to hire servants to work for her. Men have worn their share of impractical clothing, however. The late Gothic houppelande, for example, a courtly style worn by both sexes, was far too voluminous for peasants to work in, even if they could have afforded all the material necessary for its manufacture. The best illustrations of the new garment are found in *Les Très Riches Heures du duc de Berry,* at the Condé Museum, Chantilly, Fr. These show that the duke wore the houppelande down to the floor, but his servitors, who needed to move more freely, wore shorter gowns. Length thus provided an immediate signal of status.

The foregoing discussion does not attempt to be a comprehensive introduction to even one influence on fashion; it merely tries to suggest some of the ways in which costume can be analyzed and interpreted. Similar treatments of four other factors affecting fashion follow.

DISPLAY OF THE HUMAN PHYSIQUE

Martial display

Male sexual display at its most blatant can be seen in parts of Papua New Guinea, where the men wear bamboo penis covers that are sometimes up to 15 inches long. The purpose is to impress both women and enemies, by showing that the warriors are more virile than their opponents. The competition between warriors has led to a great variety of additional adornments such as boars' tusks, animal skins, animal teeth, claws, feathers, shells, metal pieces, bamboo, and the use of paint. In general, the more naked a society is, the more body paint is employed to denote the warriors and the chiefs, with each rank having its individual pattern. In addition, in many societies, only after an individual has reached a certain age or satisfied some other requirements is he allowed to wear certain colours or decorations. Sometimes each item of adornment represents a specific achievement, so that the more decorations a man wears, the better, braver, or more powerful he is shown to be.

Such martial display in Europe reached its apex with the tournaments of the Middle Ages. The males spent fortunes on enameled armour, ostrich plumes, pearl-embroidered tabards, ornate saddles and horsecloths, fine mounts, their retinue of grooms and squires, weapons, tents, and their declamations or speeches. It was a formalized kind of warfare, and foreign ambassadors were invited to be impressed by the martial display of the king or prince. An audience of females was also essential, as they had to confer favours on the knights, and the lady of the tournament had to present the bejeweled prize to the overall victor.

Such blatant display as bamboo penis cases was typified in Europe by the codpiece. During the 14th century men started shortening their tunics until they reached the crotch. A special pouch, the codpiece, had to be created to fill in the gap between the hose at the top. Initially the codpiece was not padded, but it grew larger until by the 1540s the Spanish were wearing a vertical, or erect, codpiece. This style—and its spread to other parts of Europe—may be seen to be a reflection of Spain's new dominance in the Western world and its new wealth. Spanish pride and influence were manifested in vertical codpieces, but they were soon deflated by England's Queen Elizabeth I and her navy. Perhaps in recognition of the arrival of queens regnant in England and Scotland, as well as a queen mother regent in France, both men's and women's dress began to feature a more rounded, "feminine" silhouette, and codpieces began to be covered up. Soon, female width, in the shape of the farthingale, caused codpieces to disappear completely, as men's breeches were padded out to match the ladies' skirts.

A covered-up look then dominated male attire until the late 18th century, when the Neoclassical movement led to tighter, more revealing clothes. Skin-coloured knee breeches in buckskin became the rage, and waistcoats shrank, so that from the waist downward the male form was again on show. A naked style affected the army too; uniforms became skintight, and the male form was displayed most obviously in the Napoleonic period. Under Queen Victoria the frock coat concealed all such shocking elements as legs, waist, and bulge, which remained concealed until after World War II, when skintight jeans became the means for a renewal of male sexual display. By the 1990s, Lycra (trademark) had entered at least some men's wardrobes in the form of leisure wear, its clinging characteristics providing even more extreme "naked" outlines. Thus, since the 14th century in the West, the degree of exposure of the male body has alternated between total concealment and complete display.

Views on female display have also changed dramatically. In primitive societies wives were often loaded with copper necklaces, earrings and bangles to display the wealth of their husbands. Until the 20th century, Bulgarian peasants similarly bedecked the women in their families with coins, disks, pieces of mirror, and chains to show their economic status. Such practices may date back to ancient Greece, where it was the custom in Athens to dress statues of female goddesses with new clothes and jewels every year.

Modesty in women's dress

In antiquity total nudity was acceptable for both sexes in gymnasia, at funerals, and in temples. It was only with the rise of Christianity, and 600 years later Islām, that modest covering of the female form became compulsory. St. Paul wrote to Timothy that women should not display, "that women should adorn themselves modestly and sensibly in seemly apparel, not with braided hair or gold or pearls or costly attire but by good deeds, as befits women who profess religion." St. Peter expressed similar views, and St. Augustine of Hippo went even further by censuring makeup as well, although he allowed that a woman might adorn herself slightly to please her husband if the practice was carried out in private.

Once Theodosius I made Christianity compulsory in the Roman Empire in 381, Christian views on modesty dominated women's appearance, with the exception of the imperial court; imperial princesses—like the emperor—were permitted to continue decorating themselves in luxurious finery. Some Church Fathers and churchmen expected

Photograph, © Reunion des Musees Nationaux



Figure 41: The impractical aspect of court fashion, as seen in the long sleeves and trailing gowns of 15th-century courtiers. The servants wear shorter gowns, an immediate sign of their status. Detail from "A Fishing Party at the Court of Holland," gouache by an unknown Dutch or Flemish artist, *c.* 1425. In the Louvre, Paris.

fashions to cease to change at this point, but this was an unrealistic attitude from the first. The slow changes that did occur, however, did little to alter the modest style of women's clothing. Then Diocletian divided the Roman Empire into two parts, and Constantine I the Great founded another capital, Constantinople. Inevitably, each centre felt the need to establish its own identity in clothes and styles. Eastern Rome on the Bosporus adopted the Eastern taste for coloured and patterned fabrics, and in 552, when the emperor Justinian established the first silk-manufacturing industry in Europe in Constantinople, the city became renowned for its luxurious silks and brocades.

Meanwhile, western Rome suffered barbarian invasions and centuries of disorder, until it broke up into separate kingdoms. Once these new courts had established themselves, it was only a matter of time before they, too, started trying to outdress and outshine one another. The Anglo-Saxons, for example, wore loose clothes, but after the Norman Conquest a change followed. By the 1090s members of the Norman court had started wearing tighter-fitting clothes. This was achieved by cutting the garments on the bias and lacing them under the arm, with the result that the female figure in particular was outlined very obviously. Although abbots and bishops objected vehemently, the new fashion for displaying the physique continued unabashed. This style, which also featured exaggerated cuffs reaching the ground and waistlines down at the hips, dominated the 1100s. A looser gown with a normal waistline eventually ousted the fashion in the 1200s. This pleased the church, but the desire for change in the West was too well founded to fade away. By the 1340s exposure was back, this time with necklines so wide that they were almost off the shoulder. Moreover, the adoption of buttonholes from the Moors around 1250 had introduced the art of tailoring. Clothes could now be cut very tight and still be easily removed. Shaped seams evolved, waistlines dropped again, and the possession of a shapely figure was essential for both men and women (although men were permitted to use padding over the chest to give them a curved frontage). When Edward III defeated the French at Crécy in 1346, the chronicles said it was God's punishment on the French court for wearing such shocking styles, a comment that completely overlooked the fact that the Anglo-Norman victors were wearing exactly the same fashion.

Movement toward tighter-fitting clothes

By 1364 the houppelande gown had been invented, which did conceal the human figure better, but by 1400 the church was complaining about the excessive amount of material used in the skirts and sleeves of the houppelande. Some houppelandes concealed the neck; others had a curved neckline that was almost as wide as that of the 1340s. This latter style was worn exclusively by women. Women's waistlines were higher, too, emphasizing the bosom and making the differences between the sexes obvious. The exposure of the female neck became almost permanent in court circles thenceforth. A V-shaped neckline that dipped to the bust and had to be filled in with a stomacher began to replace the wide one by the middle of the 1400s. By the 1490s the square neckline of the Tudor style, which exposed the female throat and chest from shoulder to shoulder, had become the dominant mode. When ruffs began to develop, there was a spell of concealment for the female neck, for the cartwheel ruff was built right up to the throat in the 1580s; but by 1588 Elizabeth I of England had adopted the open-fronted Medici ruff, and the exposure of the woman's throat returned as a permanent feature of court style. (Puritan ladies of course concealed the neck completely, but they tried to avoid fashion styles and trends.)

The majority of the changes in fashion that occurred from the late 1400s to the late 1700s were fueled by court competition rather than by changing attitudes about appropriate amounts of display or modesty for women. The dominant state usually affected the styles worn at other capitals during this period. It was not until the end of the 18th century, when Neoclassical taste came to the fore, that the exposure of the female form was again a major issue.

When the English novelist Fanny Burney visited Paris in April 1802, her modest wardrobe was found too full: "Three petticoats! no one wears more than one! Stays? everybody has left off even corsets! Shift-sleeves? not a soul now wears even a chemise." To be purely classical, young women adopted high-waisted, diaphanous gowns, with only one petticoat beneath; the whole of the female frame was flaunted in a manner that seemed indecent to the older generation. Stays were not abandoned for long, however, as the new slim line required a sylphlike, figure, and any bulges of the stomach or bottom had to be suppressed. In any case, this classical vogue was not strictly accurate. The gowns were not draped on or pinned together as in antiquity but had a tailored bodice and cap sleeves. In general, revivals of past styles only approximate a look and never adopt the whole vocabulary.

With the rise of Romanticism a more covered-up style developed. Women had to become demure maidens, hiding their faces in poke bonnets and concealing their figures under petticoats and shawls. By 1856 the cage crinoline of steel took this isolation of the saintly maiden to its extreme, by making her unapproachable. At this point, haute couture entered the fashion scene. The great couturier Charles Frederick Worth flattened the front of the crinoline in 1864, and in the winter of 1867–68 he abolished the garment completely in favour of long trains. In 1869 he revived the Baroque bustle but five years later slimmed the skirt down and launched the longer, fitted cuirass bodice. The narrower skirts allowed the outline of a woman's legs to be seen for the first time in 50 years. It was a brief pleasure. In 1881 Worth revived bustles, this time in a squarer, sharper look, and the exposure of the female person was ended. Only in evening dress was the bosom disclosed, and anything below that point was unseen and unmentionable.

The demure look of the Romantic period

Also influential at this time was the vogue for women's sports. There had been a similar strong fashion for women playing sports in the 1770s and '80s, when archery, shooting, lawn bowling, and riding were all permitted. The Romantics later ruled out such activities as too masculine, but by the 1860s sports were creeping back for women, and some freer clothes evolved in consequence. Amelia Bloomer's reformed trousers for women did not become fashionable, but they were adopted by women gymnasts and sea bathers. Short skirts were designed by Worth for walking, and short hems spread into golfing, shooting, and tennis outfits, while bloomers were worn for cycling. The trend to make clothes more comfortable and to reduce the amount of underwear worn was initiated by this sporting activity. When women could rush about without fuss playing croquet, rounders, and hockey during the day, they did not want to have to don elaborate confections for dinner and evening. This desire to lighten the wardrobe took time to spread through the world of fashion, however, and it was not until the next classical revival in 1907–08 that evening dresses became simpler.

The acceptance of less cumbersome costumes for sports affected swimwear, too, and, once the designer Gabrielle "Coco" Chanel made suntans the rage, exposure of the female form became almost total. Sunbathing suits revealed more of the female anatomy than any costume in history since antiquity. (Whereas in the past ladies had gone to great lengths to avoid being browned by the sun, for a sunburned complexion was the mark of a peasant, there was an almost universal vogue for sun worship in the West from the 1920s until the 1980s, when such exposure of the physique began to be warned against, with doctors stressing the dangers of skin cancer.) The backless evening dresses of the 1920s and '30s required a suntan to display and in cut were practically bathing costumes with skirts. The 1950s launched the bikini, which provided minimal coverage for women, and since then even total nudity became acceptable on some beaches.

Swimwear

GOVERNMENT REGULATION OF DRESS

Sumptuary laws. For thousands of years governments have tried to control spending by employing sumptuary laws. The first such law under the Roman Republic, the Lex Oppia, was enacted in 215 BC; it ruled that women could not wear more than half an ounce of gold upon

their persons and that their tunics should not be in different colours. Most Roman sumptuary laws tried to control spending on funerals, banquets, and festivals; there were no further laws on dress until the emperor Tiberius ruled that no silken clothing should disgrace men. Such a soft fabric as silk was considered fit only for women; the Roman male was to be a tough and severe character who did not wear Eastern imports. By 303 AD, however, Diocletian's Edict on Maximum Prices mentions the *sarcinator*, a professional tailor who made only silk clothing, and so the business seems to have expanded despite Tiberius.

It was not until the 1300s, when national governments had been established in France and England and city-states formed in Italy, that sumptuary laws appear in any number in the West. In 1322 Florence forbade the wearing of silk and scarlet cloth by its citizens outside their houses. In 1366 Perugia banned the wearing of velvet, silk, and satin within its boundaries. The impact of such legislation can be seen in the wardrobe of Francesco di Marco Datini, a merchant of Prato. Despite the fact that he had business houses from Avignon to Spain as well as in Italy and was the equivalent of a modern millionaire, his finest gowns in 1397 were made of woolen cloth, their only hint of luxury provided by a taffeta lining. The law did not permit the commercial classes to own garments made of velvet, brocade, silk, or other rich fabrics.

Dual purpose of sumptuary laws

Whereas Roman sumptuary law had applied equally to all women and all men, in western Europe the laws were more discriminatory, restricting the richest fabrics, furs, and jewels to the aristocracy. Thus, in England in 1337 Edward III ruled that no one below the rank of knight could wear fur. The same law also decreed that only English-made cloth could be worn in England. This dual role of ensuring class distinctions and banning imported goods was common in sumptuary law. In 1362 Edward III issued another edict aimed at preventing people from dressing above their station. Merchants could wear the same clothes as an esquire or knight, but only if they were five times wealthier. Yeomen and below could not wear silk, cloth of silver, chains, jewels, or buttons (which were then made of expensive materials or gems). They were not to wear the short coats or tunics worn by noblemen. Carters, plowmen, shepherds, oxherds, cowherds, swineherds, dairymen, and farm labourers were to wear only russet cloth at a shilling a yard and undyed blanket cloth. Thus, farming folk were restricted to natural wool tone and russet, and they continued wearing such colours into the 20th century. Only lords might wear cloth of gold and sable furs. Esquires and gentlemen were not allowed velvet, satin, ermines, or satin damask unless they were sergeants of the royal household. Women could not wear gold or silver girdles, nor foreign silk head covers.

Similar laws explicitly stipulating the fabrics, styles, and colours to be worn by men and women of particular social or economic standing were issued in Spain and France as well. Furthermore, in France and England it was often claimed that such laws were issued for moral or religious reasons. For example, in 1583 Henry III of France decreed that in order to regularize and reform clothing, which was dissolute and superfluous, the wearing of precious stones and pearls on garments was restricted to princes. The richest fabrics allowed were velvet, satin, damask, and taffeta, all without any enrichment beyond silk linings. Bands of embroidery in gold and silver were banned. Henry III stressed that God was angry because he could not recognize a person's quality from his clothes. A similar excuse had been given in England in 1463 when Edward IV issued a sumptuary law on the grounds that God was displeased by excessive and inordinate apparel.

In the 17th century sumptuary laws were increasingly used to restrict foreign imports and had less to do with status than with trade wars. France, for example, was trying to set up its own silk industry and therefore banned Italian silks and English cloth. Italy and Spain, however, continued issuing class restrictions on dress until 1800.

Other types of legislation. In Russia clothing law was used to modernize the country. As soon as Tsar Peter I the Great returned from working in the dockyards of Amsterdam and London in 1697–98, he began requiring his princes to shave their beards. Then in 1701 he ruled that his subjects must adopt Western dress. Peter's command applied to both men and women but at first affected only members of the court and government officials. Merchants and peasants continued to wear traditional garments into the 19th and sometimes even the 20th century.

Use of clothing laws to encourage modernization

A similar attempt to modernize a nation through its clothing was made by Mustafa Kemal (known as Atatürk) in Turkey in 1925. Laws were passed banning the fez and requiring Panama hats to be worn. To some Turks, wearing Western attire instead of traditional garments was akin to heresy, but Mustafa Kemal succeeded in changing dress, in the cities at least. With the rise of fundamentalist Islām in the late 20th century, Western styles of dress again became a subject of controversy in Turkey. Some Turks demanded that women be required to cover their heads and men to wear beards. The government responded by imposing fines on women who wore head scarves as a Muslim gesture.

In other countries, clothing legislation has been passed to ensure the preservation of local identity and dress in the face of encroaching foreign cultures. In Iran, for example, following the Islāmic revolution in the late 1970s, laws that had encouraged Western customs and clothing were replaced by ones that enforced traditional Islāmic codes of dress and behaviour.

In the West the most recent government restrictions of clothing occurred during World Wars I and II, when shortages prompted the establishment of clothes-rationing systems.

REBELLION

Rebellion against the established or dominant fashion has been a constant theme in the history of costume. The reasons prompting such rebellion are various: to shock, to attract attention, to protest against the traditional social order, to avoid current trends and thereby avoid dating oneself. One of the earliest forms such rebellion has taken—and continues to take—has been that of women adopting male dress. By donning men's clothing, women have been able to challenge the status quo and participate in activities or roles traditionally perceived as masculine.

Women's adoption of male dress

There are several examples of women in antiquity who put on male armour to go to war. Herodotus cites Queen Tomyris of the Massagetai, who led her troops against Cyrus II the Great of Persia and killed him in 529 BC. Herodotus also records Queen Artemisia I, admiral of her own ships in 480 BC when she sailed with the navy of Xerxes I, who valued her opinions highly. Queen Boudicca of the Iceni tried to drive the Romans out of Britain in AD 61. The Saxon King Alfred appointed his daughter Aethelflaed commander in chief of the west, and she successfully liberated Derby and Leicester from the Danes in 917–918. In 1080 Duchess Gaita of Lombardy rode in full male armour alongside her husband. Princess Anna Comnena of Constantinople called Gaita a "formidable sight."

The practice of women wearing male dress has not always been accepted, however. In 1429 Joan of Arc adopted male clothes, and this wearing of male dress was included among the charges against her when she was tried by the bishop of Beauvais. The bishop said her claim that God, angels, and saints had told her to don male attire was contrary to the modesty of women, was prohibited by divine law, and was forbidden by ecclesiastical censure on pain of anathema. If her voices had told her to dress as a man, why had she chosen such short, tight, and dissolute garments as tabards, cottes, and elaborate hats, and why had she cut her hair like a man, with a shaved neck? Joan confessed to error and was ordered to wear women's clothing. Nevertheless, she reverted to male dress in prison, which the bishop claimed was a sign that she had reneged on her confession. On further questioning, Joan recanted her confession and was condemned to be burned.

It has not been only for reasons of war or to defend their homes that women have adopted men's clothing. British historian Henry Knighton complained in 1348 that some 40 or 50 English ladies were arriving at tournaments in male dress and armour to parade in the intervals, so that they might share in the glory of a tourney. Knighton



Figure 42: Madame de Saint-Baslmont, a 17th-century French noblewoman who adopted men's clothing and armour to defend her estates during the Thirty Years' War. "Madame de Saint-Baslmont on Horseback," oil on canvas by Claude Deruet, *c.* 1640. In the Musée Historique Lorrain, Nancy, France.
Musee Historique Lorrain, Nancy; photograph, Gilbert Mangin

claimed that God so was incensed at this behaviour that he sent thunderstorms to drive the women indoors.

Women also have found men's clothing more suitable for certain types of work. The women pirates Mary Read and Ann Bonney donned male trousers when at sea until their capture in 1720. In 1745 Britain's Hannah Snell joined the marines and served in India for five years, wearing a male uniform all the time. It was not only a wish for action that made some women adopt male clothing. In the 19th century there were several examples of women doing so in order to earn a man's wages, which were higher than a woman's. In 1818 Helen Oliver in Scotland met a plowman who turned out to be a woman, so she copied the idea and, borrowing her brother's suit, went off to work as a plasterer. By 1866 Helen Bruce had been working in male dress since she was 17, as an errand boy, shop lad, ship's stoker, tallyman at a mine, and clerk. As women were not allowed to become doctors, Miranda Barry dressed as a man and obtained a degree in medicine at the University of Edinburgh. She then became an army surgeon and ended her career as inspector general of military hospitals in Canada in 1857, after serving in the Crimean War.

Rebellion against contemporary fashions

Cultural rebels have often chosen to adopt antique fashions in order to reject, or at least distance themselves from, their own time or to identify with what they believed to be a superior age. Sometimes such borrowings from the past become a widely accepted fashion, as in the late 18th and early 19th centuries, when Neoclassicism was at its height and women's gowns were supposed to be based on ancient Greek and Roman styles. More frequently, however, the practice remains on fashion's fringes. It has nevertheless persisted since ancient times.

The Roman empress Messalina Valeria led a revolt against Roman dress by wearing Greek clothes herself (coloured Ionic chitons fastened down the arms with bejeweled brooches) and by wearing her hair in Greek hairnets and tiaras. Her male friends similarly wore coloured Greek cloaks instead of the chalky white Roman toga. More recently, in the 1960s and '70s, many young men and women in the United States adopted the "granny" look. By wearing garments that had been popular 100 years before, such as collarless shirts, long, high-waisted cotton dresses, and small, metal-rimmed "granny" glasses, the wearers expressed their disdain for the contemporary adult establishment and their dress.

Artists have similarly often preferred older fashions, but this is usually because they wish to achieve an effect of timelessness. Leonardo da Vinci wrote in his *Treatise on Painting*, published long after his death, that art should avoid the fashion:

> As far as possible avoid the costumes of your own day. . . . Costumes of our period should not be depicted unless it be on tombstones, so that we may be spared being laughed at by our successors for the mad fashions of men and leave behind only things that may be admired for their dignity and beauty."

He showed how to tackle the problem in his portrait of "Mona Lisa" (Louvre, Paris), by dressing her in a coloured shift, loosely pleated at the neck, instead of the tight clothes that were then popular. This concept spread through western Europe over the following centuries. In the 17th century many rulers were depicted as Roman emperors in Roman armour, considered the ideal symbol for the age of absolute monarchy, and it became a sign of sophistication to look Roman in one's portrait, even if the sitter was wearing a periwig at the same time. (People were reluctant to change their hairstyles to an antique manner, as they had to wear them outside the artists' studios.) In the 18th century, aristocrats had copies made of the clothes in their ancestors' portraits to wear at masquerades and in their own portraits. Although the practice was a cultural revolt against the tyranny of contemporary fashion, the clothing was generally expressed with current tastes in mind. Artistic reform of dress in the 19th century was initiated by the Pre-Raphaelites in 1848, and by the 1860s the "Aesthetic" dress they promoted began to be adopted in sophisticated societies. The invention and widespread use of photography has effectively abolished any further need for the establishment of a specific clothing policy for art in opposition to that of high fashion. It has become acceptable for painters and sculptors—like photographers—to render contemporary fashions accurately. Extreme trends are still usually avoided, however, and portraitists of royalty often use uniforms and robes of orders of knighthood to confer a historical character.

The use of underwear as outerwear

The desire to shock has led to many rebellions in fashion. In 1783 Queen Marie-Antoinette introduced her revolutionary white muslin chemise dress—to the horror of the French silk industry. It scandalized the elderly and the conservative, who considered the chemise an undergarment, but it changed the prevailing mode of dress from one featuring hoops and brocades to the Neoclassical style emphasizing plain white muslin and the natural figure. Such use of underwear as outerwear has been recurrent in fashion history and has continued into modern times, as

By courtesy of the National Gallery of Art, Washington: Timken Collection



Figure 43: Queen Marie-Antoinette wearing the revolutionary white muslin dress that shocked conservative society. "Queen Marie-Antoinette," oil on canvas, attributed to Élisabeth Vigée-Lebrun, *c.* 1783. In the National Gallery of Art, Washington, D.C.

can be seen by the popularity of the bustier among young women in the 1980s.

Similarly, the desire to shock has remained a constant, especially among the young, who since World War II have had a significant influence on the fashion scene. Postwar teenagers have had both the money and the leisure time necessary to reject the established order and to devise a look of their own. Included among the styles they introduced are the T-shirts and jeans of the 1950s, the long-haired hippie look of the 1960s, and the punk look of the late 1970s.

EXOTICA

Like rebellion, the adoption of foreign elements has been a constant theme in the history of dress, and it too dates to antiquity. The first exotic fabric to reach the West was silk from China, which the Persians introduced to the Greeks and Romans and which has remained popular to the present. Another early import was the caftan coat, which is believed to have originated in Central Asia and which appeared among the Hittites, the Assyrians, and the Medes and Persians by 700 BC. During the Hellenistic period Greek tunics were introduced into the Middle East, but the caftan continued to be worn in Persia. The caftan eventually made its way to Russia, where it was described by the Arab traveler Ibn Fadlan in AD 922 when he saw a Viking chief's funeral on the Volga; the chief's body was dressed in a caftan of cloth of gold with golden buttons and a gold cap trimmed with sable. The Turks also adopted caftans, and they then brought the style to Hungary and Poland when they conquered those lands. Subsequently, there were occasional vogues for Turkish dress in Italy, Germany, and England, and the caftan became the model for later Western garments featuring fitted backs and open fronts.



Figure 44: Silk Indian jacket and pajamas worn by an English earl. Detail from "William Feilding, 1st Earl of Denbigh," oil on canvas by Anthony Van Dyck, *c.* 1633. In the National Gallery, London.
By courtesy of the National Gallery, London

Introduction of kimono and pajamas

The Japanese kimono entered the Western wardrobe in the 17th century. The English called the garments Indian gowns, probably because the East India Company imported them, but the Dutch more accurately called them Japanese coats. The garment was also termed a nightgown and a banyan and became fashionable for undress. The diarist Samuel Pepys bought himself an Indian gown on July 1, 1661, for 34 shillings. He further recorded that on Nov. 21, 1666, "I to wait on Sir Ph. Howard, whom I find dressing himself in his night-gown and Turban like a Turke." Strictly speaking, the Indian gown was meant to be worn for informal, private occasions, but a superior like Sir Philip Howard could wear such clothing to receive underlings, though they had to be fully dressed to attend him. The first nightgowns were cut loose like the Japanese originals, but in the late 18th century they became more fitted and tailored like coats. Such dressing gowns have remained fashionable and are now known as housecoats, bathrobes, wraps, and negligees depending on the material used. Indian pajamas, a soft cotton suit consisting of trousers and a loose, fitted jacket fastened down the front, were also introduced into Europe in the early 17th century. They, too, have remained popular for undress, although the style has sometimes also been adopted for more formal wear.

Many foreign garments are copied or borrowed of necessity. For example, when the Europeans invaded the Americas, the English and the French were quick to adopt Native American moccasins because few of the settlers knew how to make shoes. Similarly, in Canada the Indian snowshoe was essential wear, and many hunters and trappers adopted Indian fringe on their deerskin tunics as a practical embellishment that allowed rain to run off. When winter sports became fashionable in the 20th century, Eskimo padded boots and parkas (hooded jackets) were copied.

The modern Western wardrobe can include elements of Asian, African, and Native American dress. Similarly, non-Western cultures have adopted some Western garments, particularly the Western-style suit for business wear. In the future, as improved transportation and communication technology effectively shrink the size of the world, foreign influences on dress will no doubt continue to be introduced with increasing speed and influence.

(D.J.A. de M.)

JEWELRY

Jewelry consists of objects of personal adornment prized for the craftsmanship that went into their creation and generally for the value of their components as well. Through the centuries and from culture to culture, the materials considered rare and beautiful have ranged from shells, bones, pebbles, tusks, claws, and wood to so-called precious metals, precious and semiprecious stones, pearls, corals, enamels, vitreous pastes, and ceramics. In certain eras artist-craftsmen have sometimes placed less emphasis on the intrinsic value of materials than on their aesthetic function as components contributing to the effect of the whole. Thus, they might fashion a brooch out of steel or plastic rather than gold or platinum. Furthermore, in addition to its decorative function, during much of its history jewelry has also been worn as a sign of social rank—forbidden by sumptuary laws to all but the ruling classes—and as a talisman to avert evil and bring good luck. During the Middle Ages, for example, a ruby ring was thought to bring its owner lands and titles, to bestow virtue, to protect against seduction, and to prevent effervescence in water—but only if worn on the left hand.

Materials and methods

The first materials used to make objects for personal adornment were taken from the animal and vegetable world. The material taken from the animal world, in a natural or processed form, constituted the actual adornment, whereas vegetable fibres served as its support. A great variety of shells and pieces of shell were used during the prehistoric age and are still used in certain island and coastal cultures to make necklaces, bracelets, pendants, and headdresses. In the inland regions the first materials used for personal adornment came from mammoths' tusks, the horns of reindeer and other animals, and, later on, amber and lignite.

All materials that have been used over the centuries for the manufacture of jewelry have undergone to some extent mechanical, physical, or chemical treatment for the purpose of transforming their raw shapes into shapes that, in addition to being functional, also satisfy certain aesthetic concepts.

METALS

Gold, silver, and platinum

Precious metals and their properties. Of gold's properties, when it was first discovered (probably in Mesopotamia before 3000 BC), it was the metal's malleability that was a new phenomenon: only beeswax, when heated to a certain temperature, could be compared to it. Gold's molecules move and change position in accordance with the stresses to which it is submitted, so that when it is beaten it gains in surface area what it loses in thickness. In modern jewelry, gold can take on a variety of hues when it is alloyed with other metals: water green, white, gray, red, and blue.

After gold, silver is the metal most widely used in jewelry and the most malleable. Although known during the Copper Age, silver made only rare appearances in jewelry before the classical age. In general, silver was, and still is, used in jewelry for economic reasons or to obtain chromatic effects. It was often used in the 17th, 18th, and 19th centuries, however, as support in settings for diamonds and other transparent precious stones, in order to encourage the reflection of light.

Another rare metal, whose use in jewelry is fairly recent, is platinum. From the 19th century onward this metal was used ever more frequently in jewelry because of its white brilliance and malleability, as well as its resistance to acids and its high melting point.

Modern jewelry, such as that designed by early 20th-century artists, introduced nonprecious metals such as steel.

Sheet metal, embossing, openwork, and granulation

Metalwork. The basic components of jewelry have always consisted of sheet metal, metal cast in a mold, and wire (more or less heavy or fine). These components take on the desired shape by means of techniques carried out with the help of tools. Gold in its natural state was beaten while hot or cold and reduced to extremely thin sheets (this operation could be performed with stone hammers). The sheets were then cut into the desired sizes.

Examination of the most ancient pieces of jewelry shows that one of the techniques used most widely in decorating metal sheets for jewelry was embossing (relief work). Throughout the centuries embossing techniques have remained substantially unchanged, although in modern times mechanization has made possible mass production of decorative parts of jewelry, with great savings of time and labour but with a corresponding lack of art.

In repoussé the relief is pressed (in a negative mold) or hammered out from the reverse side of the gold sheet and then finished off on the right side with a hammer or engraving tool. For half-modeled or completely round reliefs, the gold leaf was pressed onto wooden or bronze models. Completely round objects were made in two pieces and then welded together.

Another embossing, or relief, technique is engraving, which involves impressing designs into the metal with a sharp tool.

Decorative openwork designs can be created by piercing the gold leaf. In the Roman period this technique was called *opus interassile*.

Granulation is a decorative technique in which small or minute gold balls (with diameters ranging from $1/60$ to $1/180$ of an inch) are used to form silhouettes on smooth or embossed metal.

Casting from precious metals has always been rare. When the relief was to be visible only from one side, the metal was poured into the cast and, when hardened, touched up with a graver. When the relief was to be fully modeled, the cire perdue (lost-wax) process, involving casting from a wax mold, was used.

Metal wire and filigree

Gold and silver wire, according to its function, can be made into various sizes, shapes, sections, and weights. It can serve to join, to support pendants of varying importance, to make necklaces and bracelets, or to alternate with other decorative components.

From the 3rd millennium BC through the present day, chains—ranging from the simple type, consisting of a series of round or oval rings, to one of the oldest elaborations, the "loop in loop," or square, chain—have offered goldsmiths the widest field for decorative imagination.

Filigree is a form of decoration made exclusively from fine gold or silver wire welded onto the surface of an object made of the same metal or done in openwork (without a background). The decoration to be carried out is designed first on a model with a flat or curved surface identical to that on which the completed filigree is to be welded or to the unsupported shape that it must assume. It can be made from smooth wire or from a ropelike plait or from a series of small hemispheres. A more complicated type of filigree consists of metal wire made in the shape of beads called granulated filigree.

After having been prepared separately, the different parts that make up a piece of jewelry are put together. In primitive jewelry this was done mechanically, by inserting beaten pins, by bending and beating the parts to be fastened together, or by binding them with gold wire or tape. Welding is a technique belonging to a more highly developed stage of ancient goldworking (end of the 3rd millennium BC).

Enamel work. In enamel work, powdered glass coloured with metal oxides diluted with water and adhesive is applied to certain parts of the piece of jewelry that have been cut lower or surrounded with a raised rim made of gold, silver, or copper. The object is then heated until the glass melts and adheres to the metal. As the enamel gradually cools, it crystallizes and, when smoothed, takes on greater lustre and colour. The enamel applied to jewelry can be opaque or translucent. By letting light through, transparent enamel catches reflections from the metal to which it is applied and makes visible any engraving done on the metal. Enamel is also distinguished according to the way it is applied, as in cloisonné, champlevé, basse taille, painted, and *plique à jour*.

Enameling preceded the polychromy created by precious stones. In the beginning, in Egypt, Greece, and the Sāsānian period in Iran, unpolished enameled parts of jewelry were often used to imitate lapis lazuli or malachite.

Niello technique

To a limited extent, jewelry also was decorated with the niello technique (from the Latin *nigellus*, an adjective derived from *niger*, meaning black). This consists of cutting grooves in gold or silver with a graver and then filling these with a powder made of red copper, silver, lead, sulfur, and borax. When heated, the powder melts and fills the grooves, adhering to the metal. After the piece has cooled, the surface is smoothed and polished, and the design shows up in black.

GEMS

In addition to gold, silver, and platinum, the precious materials most widely used in jewelry are gems—any precious or semiprecious stone. By definition this group also includes some animal and vegetable products with precious characteristics, such as amber, pearls, and coral. Conventionally, the following are classified as precious stones: diamonds, rubies (corundum), emeralds (beryl), and sapphires (corundum). To these, however, can be added chrysoberyl, topaz, and zircon because of their hardness and their refraction and transparency index.

Use of diamonds, rubies, and sapphires

The properties of gems. Diamonds have the highest refraction index, and those used for jewels are very transparent. Diamonds from Indian deposits were known in ancient times; in the West the limited use of diamonds began in the late Middle Ages. Diamonds for jewelry are graded on the basis of colour from blue-white to yellow. Grading also is done on the basis of purity, which varies from perfectly clear, extremely pure stones to those with many impurities and flaws. Large demand provided an incentive for the production of false diamonds (as well as other stones) as early as 1675 in Paris.

Mogok rubies, from Myanmar (Burma), are the most highly prized because of their bright red colour (pigeon blood). Those from Thailand are usually a more brownish colour, while those from Sri Lanka tend toward violet. Production of synthetic stones is far greater than the sup-

ply of natural rubies. The physical and optical properties of synthetic and natural rubies are so similar that it is difficult to distinguish between them.

The sapphire (blue variety of corundum) is considered one of the most valuable of precious stones. A sapphire's colouring usually indicates its origin. Those from Myanmar are deep blue. The Kashmir (Indian) sapphire is cornflower blue and is highly prized, being quite rare. Sapphires from Thailand are very similar in colour to those from Myanmar; those from Sri Lanka are of different shades but incline toward violet. Sapphires, like rubies, can be cut so that, in the light, a beautiful, luminous six-pointed star appears on the surface of the gem. Star sapphires and rubies are semi-opalescent. Synthetic sapphires and rubies are produced by the same industries.

Use of emeralds

The green emerald is a precious stone used since very ancient times. There is documentation of its presence in Egypt during the life of Pharaoh Sesostris III in the 19th century BC. At the end of the 16th century, emeralds from South America were brought into Europe. On the American continent, the first peoples to use emeralds were those belonging to the pre-Columbian civilizations, in particular the Inca. In 1935 in the United States (Chatham) and in Germany (Farbenindustrie), synthetic emerald crystals were made with characteristics similar to natural ones.

Among the beryls, mention must be made of morganite (pink beryl) found in various shades of peach-blossom pink. The main deposits are in California and Madagascar.

The two best known and most widely used varieties of chrysoberyl are alexandrite (transparent) and Oriental cat's-eye (opaque). Because of its great power of absorption of certain colours, alexandrite looks green in daylight and reddish purple in artificial light. The cat's-eye is a yellowish green colour and is characterized by a luminous line. The intensity of the light in this line varies according to the brightness of the rays of light that strike it.

One of the most important gems with pure crystals is the topaz, used a great deal in jewelry. The honey-yellow variety is the best known, but there are also pink, red, blue, and the less-used colourless stones. The Oriental topaz (a corundum) and citrine quartz are also widely used. They are less rare than other kinds of topaz and, therefore, less expensive but create a similar effect.

Among the less-important and less-rare stones, the zircon is quite widely used in its three varieties: orange, blue, and colourless. The orange variety is called jacinth and was used to a great extent in classical antiquity. The blue variety is called starlite or Siam zircon, while the third type is called Ceylon or Matara diamond.

Among the semiprecious stones used in jewelry are amethyst, garnet, aquamarine, amber, jade, turquoise, opal, lapis lazuli, and malachite. Matrix jewelry is cut from a stone such as opal or turquoise and the surrounding natural material, or matrix.

Varieties of pearls

The pearl is one of the oldest gems known. Its colour varies according to the waters from which it comes. Pearls from the Persian Gulf are usually cream-coloured; those from Australia are white with greenish or bluish shades; golden-brown pearls come from the Gulf of Panama; those from Mexico are black or reddish brown; pink pearls are from Sri Lanka; and those from Japan are cream-coloured or white with greenish tones. The main characteristic of the pearl is its iridescence. Baroque pearls are those with defects in their outer layer. In modern times baroque pearls are rounded off artificially but, in the 16th and 17th centuries, their irregular form was exploited in jewelry by using them to make up parts of animals or other figures. After huge quantities of cultivated pearls invaded the world market, interest in natural pearls underwent a considerable decrease.

In addition to pearl, a number of other organic materials, including amber, coral, ivory, and jet, are considered gems.

Amber is a fossil resin, usually yellowish brown, but on occasion deep brown to red, green, or blue. It is an amorphous hydrocarbon and may contain particles of various foreign materials, trapped insects, and air bubbles. Its lustre is greasy to resinous. The most noted occurrence of amber is along the shores of the Baltic Sea, where pieces have been washed up by wave action. Other important occurrences are along the coast of Sicily, in Romania, and in Myanmar near Myitkyinā.

Coral is the skeletal material of calcium carbonate built up by small animals that live in colonies in the sea. This material is usually branchlike and occurs in a variety of colours, of which the most sought after are rose red to red. The best coral comes from the Mediterranean Sea, particularly off the coasts of Algeria and Tunisia. A black horny coral growth, probably conchiolin, which hardens on exposure to air, has been obtained off the islands of Hawaii. Coral is carved into art objects and cut as beads, cameos, and other ornaments.

The use of ivory for ornamental purposes dates to prehistory. The term should be restricted to the material derived from the tusks of certain animals—namely, the elephant, hippopotamus, warthog, walrus, sperm whale, narwhal, and the extinct mammoth (fossil ivory). The pale cream colour of new ivory darkens with age to yellow. All types are brittle and will not peel as do the plastics used to simulate them.

Jet is a dense variety of lignite formed by the submersion of driftwood in the mud of the seafloor. It has been recovered since Roman times from the shales near Whitby in northeastern England. It takes a high polish and was once popular as mourning and ecclesiastical jewelry but has been superseded by black onyx, black tourmaline, and plastics. Because it is actually a variety of coal, it will burn.

Gem engraving, setting, and cutting. The most ancient technique of stone engraving, intaglio-incised carving, was probably first used to produce seals. The art is believed to have originated in southern Mesopotamia and was highly developed by the 4th millennium BC. During the Hellenistic Age (*c.* 323–30 BC) intaglio surface engraving gave rise to the idea of carving stones in relief, exploiting the different coloured layers of certain minerals to create contrasting figures (cameo): the background was cut down to the lower level, of a different colour or shade, in order to make the subject stand out chromatically. The stones that have properties suited for this purpose are sardonyx, agate, and onyx.

Intaglio carving

The cameo is usually one of the components for necklaces, bracelets, and rings or is included in medallions with a jeweled frame. The art of cameo in jewelry was most highly developed during three periods: the late republican to early imperial period in Rome, the Renaissance, and the Neoclassical period in the 18th century.

The evolution of techniques of setting has followed that of stonecutting. The insertion of gems in jewelry can be done in various ways. The setting can have a round, square, oval, or rectangular collet (rim); in periods in which gems were mounted in their own irregular shapes, the collet followed this form. Usually, on the inside of the collet a short distance from the edge, there is a protrusion on which the stone rests. The edge is pounded down around the gem to ensure its stability. In coronet settings the form may be conical or pyramidal, solid or pierced. The edge is first shaped into a row of teeth, which are later hammered down onto the gem in order to hold it in place. Until fairly recently, nearly all gems were mounted on a metal base; and transparent stones, according to their colour, were placed on a gold or silver base to increase the amount of light reflected. As new cuts were developed for stones, setting techniques also progressed, especially for those jewels in which important stones like diamonds, emeralds, and rubies form the main theme. The tendency was to leave the stones as visible as possible (especially in rivière necklaces and bracelets made only of diamonds) by mounting them with a very small ring of white gold or platinum fitted closely against the back of the stone. Three claws, attached to this ring, hold the stone in place.

Pearls, like some coloured stones, in ancient classical times were pierced with a drill, the hole going half or all the way through according to whether the pearls were to be strung on a necklace or fastened onto a jewel.

Until the 15th century, stones were only polished or the part to be left visible was rounded into a dome shape called cabochon. The cutting known as faceting gradually developed from the first attempts in the 15th century, probably in France and the Netherlands. During the 16th

Gem cutting

century the simple rose cut began to be used, after which there were no new developments until 1640, when, under the patronage of Jules Cardinal Mazarin, the first brilliant cut was carried out (also called the Mazarin cut). Toward the end of that century, a Venitian succeeded in obtaining the triple brilliant cut, which is still used. The numerous cuts used for diamonds today are usually applied to other precious and semiprecious transparent stones as well. For emeralds, rubies, and other coloured stones the square or rectangular cut with a stepped bulb or the cabochon cut are usually used.

The history of jewelry design

The possibility of tracing jewelry's historic itinerary derives primarily from the custom, beginning with the most remote civilizations, of burying the dead with their richest garments and ornaments. Plastic and pictorial iconography—painting, sculpture, mosaic—also offer abundant testimony to the jewelry worn in various eras.

Prehistoric era

It is probable that prehistoric humans thought of decorating the body before they thought of making use of anything that could suggest clothing. Before precious metals were discovered, people who lived along the seashore decorated themselves with a great variety of shells, fishbones, fish teeth, and coloured pebbles. People who lived inland used as ornaments materials from the animals they had killed for food: reindeer antlers, mammoth tusks, and all kinds of animal bones. After they had been transformed from their natural state into various elaborate forms, these materials, together with animal skins and bird feathers, provided sufficient decoration.

This era was followed by one that saw a transition from a nomadic life to a settled social order and the subsequent birth of the most ancient civilizations. Most peoples settled along the banks of large rivers, which facilitated the development of agriculture and animal husbandry. Indirectly, this also led to the discovery of virginal alluvial deposits of minerals, first among which were gold and precious stones.

Over the years the limited jewelry forms of prehistoric times multiplied until they included ornaments for every part of the body. For the head there were crowns, diadems, tiaras, hairpins, combs, earrings, nose rings, lip rings, and earplugs. For the neck and torso there were necklaces, fibulae (the ancient safety pin), brooches, pectorals (breastplates), stomachers, belts, and watch fobs. For the arms and hands armlets, bracelets, and rings were fashioned. For the thighs, legs, and feet craftsmen designed thigh bracelets, ankle bracelets, toe rings, and shoe buckles.

MIDDLE EASTERN AND WESTERN ANTIQUITY

Sumerian. The most ancient examples of jewelry are probably those found in Queen Pu-abi's tomb at Ur in Sumeria (now called Tall al-Muqayyar), dating from the 3rd millennium BC. In the crypt the upper part of the queen's body was covered with a sort of robe made of gold, silver, lapis lazuli, carnelian, agate, and chalcedony beads, the lower edge decorated with a fringed border made of small gold, carnelian, and lapis lazuli cylinders. Near her right arm were three long gold pins with lapis lazuli heads, three amulets in the shape of fish—two made of gold and one of lapis lazuli—and a fourth amulet of gold with the figures of two seated gazelles. On the queen's head were three diadems, each smaller than the one below it, fastened to a wide gold band: the first, which came down to cover the forehead, was formed of large interlocking rings, while the second and third were made of realistically designed poplar and willow leaves (Figure 45). Above the diadems were gold flowers, on drooping stems, the petals of which had blue and white decorations. On the back of the headdress was a Spanish-type comb, with teeth decorated with golden flowers. Huge golden earrings, in the shape of linked, tapered, semitubular circles, completed the decoration of the head. On the neck was a necklace with three rows of semiprecious stones interrupted in the middle by an openwork flower in a gold circle. Many rings were worn on the fingers. There were large quantities of other jewels—among them wrist and arm bracelets and



Figure 45: Sumerian gold and faience diadems from Queen Pu-abi's tomb, Ur, c. 2500 BC. In the British Museum.
By courtesy of the Trustees of the British Museum

pectorals—belonging to the handmaidens, dignitaries, and even the horses that formed part of the funeral train. As was the custom, the queen's attendants had killed themselves in the crypt after the burial ceremony.

As this description suggests, Sumerian jewelry forms, much more numerous than those of modern jewelry, represent almost every kind developed during the course of history. Nearly all technical processes also were known: welding, alloys, filigree, stonecutting, and even enameling. Sources of inspiration, aside from geometry (disks, circles, cylinders, spheres), were the animal and vegetable world; and expressive forms were based on an essential realism enriched by a moderate use of colour.

The Tutankhamen treasure

Egyptian. The sensational discovery of the tomb of the pharaoh Tutankhamen (18th dynasty; 1539–1292 BC) revealed the fabulous treasures that accompanied an Egyptian sovereign, both during his lifetime and after his death, as well as the high degree of mastery attained by Egyptian goldsmiths. This treasure is now housed in the Egyptian Museum in Cairo and represents the biggest collection of gold and jewelry in the world. The pharaoh's innermost coffin was made entirely of gold, and the mummy was covered with a huge quantity of jewels. More jewels were found in cases and boxes in the other rooms of the tomb. The diadems, necklaces, pectorals, amulets, pendants, bracelets, earrings, and rings are of superb quality and of a high degree of refinement that has rarely been surpassed or even equaled in the history of jewelry.

The ornaments in Tutankhamen's tomb are typical of all Egyptian jewelry (Figure 46, right). The perpetuation of iconographic and chromatic principles gave the jewelry of ancient Egypt—which long remained uncontaminated in spite of contact with other civilizations—a magnificent, solid homogeneity, infused and enriched by magical religious beliefs. Ornamentation is composed largely of symbols that have a precise name and meaning, with a form of expression that is closely linked to the symbology of hieroglyphic writing. The scarab, lotus flower, Isis knot, Horus eye, falcon, serpent, vulture, and sphinx are all motif symbols tied up with such religious cults as the cult of the pharaohs and the gods and the cult of the dead. In Egyptian jewelry the use of gold is predominant, and it is generally complemented by the use of the three colours of carnelian, turquoise, and lapis lazuli or of vitreous pastes imitating them. Although there was a set, fairly limited repertoire of decorative motifs in all Egyptian jewelry, the artist-craftsmen created a wide variety of compositions, based mainly on strict symmetry or, in the jewelry made of beads, on the rhythmic repetition of shapes and colours.

Forms of Egyptian jewelry

The concept of symmetry was utilized on the small pectoral or pendant (3.3 × 2.4 inches, or 8.4 × 6.1 centimetres) that belonged to Sesostris III in the 12th dynasty (1938–1756 BC); the superbly rhythmic composition is framed by an architectonic design obtained by leaving open

all of the nonfigurative part (Figure 46, left). The jewel is coloured with carnelian, turquoise, and lapis lazuli inlays, while the function of the gold separating these materials is limited to creating the design. The victorious pharaoh is represented by two lions with the plumed heads of falcons in a symmetric position in the act of trampling conquered Nubians and Libyans. Over the scene is the protective vulture of Upper Egypt with wings outspread (Egyptian Museum). These memorial or dedicatory pendants, as well as other small jewels such as earrings, bracelets, and rings, consist exclusively of symbols.

Necklace beads—generally made of gold, stones, or glazed ceramic—are cylindrical, spherical, or in the shape of spindles or disks and are nearly always used in alternating colours and forms in many rows. The necklaces have two distinct main forms. One, called menat, was the exclusive attribute of divinity and was therefore worn only by the pharaohs. Tutankhamen's menat is a long necklace composed of many rows of beads in different shapes and colours, with a pendant and with a decorated fastening that hung down behind the shoulders. The other, much more widely used throughout the whole period, was the *usekh,* which, like the vulture-shaped necklace from the tomb of Tutankhamen, also has many rows and a semicircular form.

Of the many diadems made by Egyptian artist-craftsmen, one of the earliest was discovered in a tomb dating from the 4th dynasty (*c.* 2575–*c.* 2465 BC). It consists of a gold band supported by another band made of copper, to which three decorative designs are applied. In the centre is a disk worked with embossing in the form of four lotus buds arranged radially. On the sides are two papyrus flowers linked horizontally at the base by a disk with a carnelian, while the upper line of the flowers comes together to create a kind of nest in which two long-beaked ibis crouch. The floral and animal symbology is carried out with a style that interprets and characterizes the theme.

Among the treasures discovered in the tomb of Queen Ashhotep (18th dynasty) is a typical Egyptian bracelet. It is rigid and can be opened by means of a hinge. The front part is decorated with a vulture, whose outspread wings cover the front half of the bracelet. The whole figure of the bird is inlaid with lapis lazuli, carnelian, and vitreous paste.

Foreign influence

A first sign of outside influence occurs in the 18th dynasty and consists of earrings, which are imported jewels, unknown in classical Egyptian production. Another evidence of the influence of foreign styles in some of the jewelry of the 18th dynasty is a headdress (Figure 47) that covered nearly all of the hair, made of a network of rosette-shaped gold disks forming a real fabric (New York City, Metropolitan Museum of Art). Foreign influence increased to an ever greater extent during the last dynasties and with the arrival of the Greeks. Like all other forms of artistic expression, in spite of three centuries of Ptolemaic dynasty (up through 30 BC), the great artistic tradition of Egyptian jewelry slowly died out, and the introduction first of Hellenism and then of the Romans led to the definitive decline of the most monumental cultural and artistic structure known throughout all history.



Figure 47: Gold headdress (reconstructed; originally inlaid with carnelian and glass) of one of the three queens of Thutmose III (1479–26); from Western Thebes, 18th dynasty. In the Metropolitan Museum of Art, New York City.

By courtesy of the Metropolitan Museum of Art, New York, purchased with funds given by Henry Walters and Edward S. Harkness, 1926

Aegean. The Bronze Age civilization that flourished on the Mediterranean island of Crete is known as the Minoan. Because Crete lay near the coasts of Asia, Africa, and the Greek continent and because it was the seat of prosperous ancient civilizations and a necessary point of passage along important sea-trading routes, the Minoan civilization developed a level of wealth which, beginning about 2000 BC, stimulated intense goldworking activities of high aesthetic value. From Crete this art spread out to the Cyclades, Peloponnesus, Mycenae, and other Greek island and mainland centres. Stimulated by Minoan influ-

(Left) Hirmer Fotoarchiv, Munchen, (right) Robert Harding Picture Library



Figure 46: *Egyptian jewelry of the Middle and New Kingdoms.*
(Left) Gold pectoral with semiprecious stones belonging to Sesostris III, Middle Kingdom, 12th dynasty, *c.* 1850 BC. From Dahshur, Egypt. (Right) Pectoral of gold, silver, and semiprecious stones. From the tomb of Tutankhamen, New Kingdom, 18th dynasty, *c.* 1340 BC. Both in the Egyptian Museum, Cairo.

ence, Mycenaean art flourished from the 16th to the 14th century, gradually declining at the beginning of the 1st millennium BC.

Among the techniques used in Minoan-Mycenaean goldworking were granulation and filigree, but the most widely used was the cutting and stamping of gold sheet into beads and other designs to form necklaces and diadems, as well as to decorate clothing. The kings from Period I of Mycenaean civilization (*c.* 1580–1500 BC), discovered in their burial places, wore masks of gold sheet, and scattered over their clothing were dozens of stamped gold disks. The disks reveal the rich variety of decorative motifs used by the Mycenaeans: round, rectangular, ribbon-shaped—including combinations of volutes, flowers, stylized polyps and butterflies, rosettes, birds, and sphinxes.

A pendant from a Minoan tomb at Mallia, Crete (Archaeological Museum, Iráklion, Greece), is one of the most perfect masterpieces of jewelry that has come down to us from the 17th century BC (Figure 48). The Sun's disk is covered with granulation and is held up by two bees, forming the central part of the composition. Ring bezels (tops of the rings), with relief engravings of highly animated pastoral scenes, cults, hunting, and war, are also fine. Like those of the other jewelry forms, the ornamental motifs of the necklaces are varied, including dates, pomegranates, half-moons facing each other, lotus flowers, and a hand squeezing a woman's breast. During the late Mycenaean period, earrings appeared in the shape of the head of a bull, an animal frequently represented in early gold plate.

In addition to goldworking, Minoan-Mycenaean craftsmen also excelled at engraving gemstones for seals and rings.

Dimitri

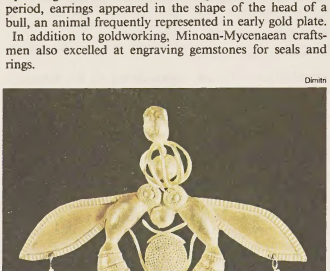


Figure 48: Minoan gold pendant of bees encircling the Sun, showing the use of granulation, from a tomb at Mallia, 17th century BC. In the Archaeological Museum, Iráklion, Crete.

Phoenician. Phoenicia, a centre for both the production and exportation of jewelry, was not a source of great originality. It is to the trading done by this people throughout the Mediterranean, however, that we owe knowledge of the products of the most highly developed civilizations in the most remote lands—northern Africa, Sardinia, Spain, and Italy. The period in the 8th and 7th centuries BC, during which Scythian-Iranian Oriental objects with their animalistic motifs were spread and consequently imitated throughout the Mediterranean countries, especially in Greece and Italy, is called the Orientalizing period.

Orientalizing period

Etruscan. In Etruria, to a much greater extent than elsewhere, the stimulus provided by the jewelry imported by the Phoenicians led to emulation that soon had imposing results. Alongside imported objects and mechanically repeated Oriental motifs, original forms, techniques, and styles developed that were the result of Etruscan taste. There was an entirely new concept, in which the goals of magnificence, impressive size, and a great wealth of decoration led to some of the most outstanding achievements in the history of jewelry. Technical virtuosity exploited all the resources available to filigree and above all to granulation, carried out with gold alone without chromatic inlaying.

Fibulae began to be made in forms other than the single Oriental leech, or boat, shape: with a dragon bow, lozenge-shaped, with a long foot. Like such ornaments as pendants and the heads of pins, fibulae were often decorated with gold dust, in which opaque granulated figures—ibexes, chimeras, sphinxes, winged lions, centaurs, horsemen, and warriors, nearly all of Oriental derivation—stand out against the smooth surface of the gold. One notable example is the fibula from the lictor's tomb in Vetulonia (Figure 49).

SCALA—Art Resource

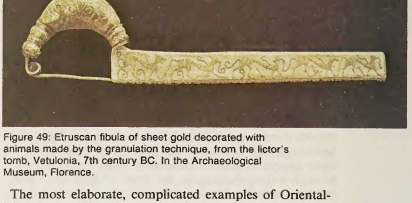


Figure 49: Etruscan fibula of sheet gold decorated with animals made by the granulation technique, from the lictor's tomb, Vetulonia, 7th century BC. In the Archaeological Museum, Florence.

The most elaborate, complicated examples of Orientalizing Etruscan jewelry consist of very large brooches with fully sculptured decoration applied to a combined tubular and plate structure. The minutely designed granulated figures of sphinxes, winged lions, chimeras, winged griffons, and human heads—set in series in alternating rows—form a plastic fabric, the details of which are of astonishing technical ability, while at the same time they suggest the evocative, mysterious animalistic symbolism of western Asian civilizations.

In the period that followed the Orientalized one, Etruscan jewelry revealed Ionic influence (6th–5th century BC). The most beautiful examples are necklaces made of many flexible chains that cross each other and bear different rows of embossed pendants in the shape of harpies, mermaids, Gorgons, and Sileni, interspersed with others such as pomegranates, acorns, lotus flowers, and palms. These show the clear influence, especially in the modeling of the pendant heads, of the Greek severe period, an influence that spread throughout the entire Etruscan territory, from Spina on the Adriatic coast of Italy to southern Italy. Even clearer evidence of the acceptance of imported forms is provided by a new shape, the bulla, a pear-shaped vessel used to hold perfume. Its surface was decorated with embossed and engraved symbolic figures.

Greek. Because gold was not readily available, jewelry was relatively rare in Archaic (*c.* 750–*c.* 500 BC) and Classical (*c.* 500–*c.* 323 BC) Greece. Examples do exist, however, and certain generalizations can be made. In the 7th and 6th centuries BC the jewelry produced in Attica and the Peloponnese shows evidence of strong Oriental stylistic influence, the same influence that in Etruscan territory turned up in a much more magnificent form. In the 5th century BC the Ionic style became predominant, taking the place of the showy Oriental style. War scenes and animals of Oriental origin disappeared, for example, from the wide oval ring bezels and were replaced exclusively by the human figure. These included naked riders on galloping horses; seated and standing maidens, depicted both with clothes and naked; and deities and mythological figures. This extremely refined repertoire in reality was more closely related to sculpture and to classic ideals of beauty than to decoration. Indeed, in its long evolution, Greek jewelry has the predominant character of sculpture in miniature and represents isolated figures or religious, mythological, or heroic scenes.

Greek expansion into Anatolia to the east, southern Italy to the west, and the Balkan Peninsula to the north

Hellenistic jewelry

resulted in the Hellenization of this entire area. Under the reign of Alexander the Great, a magnificent era for jewelry began. Hellenistic jewelry, much more so than painting and sculpture, underwent flourishing development in the art centres of the different regions under Greek rule. In the 3rd and 2nd centuries BC, the technical ability of Hellenistic goldsmiths reached the highest levels ever attained. A style both sumptuous and full of plastic vigour was created, in which meticulous arrangement of the decorative motifs resulted in the contrast and harmony, clarity and unity, rhythms and cadences that make some of these jewels complete works of art. The very fine technique and virtuosity in miniature is reflected in the creation of the first cameos and in disk earrings bearing pendants, often of minute proportions. A real masterpiece is an earring with a winged figure of a woman driving a two-horse chariot (Museum of Fine Arts, Boston). The precision of its tiny details, the severity of style with which it is modeled, and the rhythmic dynamism of the figures make this earring a microscopic monument of sculpture (Figure 50, right).

Also worthy of high consideration are the magnificent diadems that came into wide use as a result of the Persian conquests made by Alexander the Great. One type is a rigid elliptical shape with a Hercules knot in the centre and pendants hanging down over the forehead. (The Hercules knot was the most famous one used in ancient times, as it was considered a magic knot and, in jewels, took on the significance of an amulet. It also was used on bracelets, belts, and rings during this period.) Another type, decorated with jewellike enameled flowers (Figure 50, bottom), demonstrates the increasing use of colour during the Hellenistic Age.

One type of necklace that was commonly worn at this time was made of gold pieces, often hollow or filled with resin, that were fashioned into the shape of acorns, amphorae, and rosettes that sometimes alternated with stones or vitreous paste. In the 3rd century BC the bracelet in the shape of a serpent originated and remained popular through the Roman period (Figure 50, left). The serpent motif also was used for rings.

Roman. In ancient Rome, jewelry was used to an extent never seen before and not to be seen again until the Renaissance. Imperial Rome became a centre for goldsmiths' workshops. Together with the precious stones and metals that were brought to the city came lapidaries and goldsmiths from Greece and the Oriental provinces. The gold ring, which under the republic had been a sign of distinction worn by ambassadors, noblemen, and senators, gradually began to appear on the fingers of persons of lower social rank until it became common even among

By courtesy of (left) the Schmuckmuseum Pforzheim, Germany, (right) the Museum of Fine Arts, Boston; photograph (bottom) Hirmer Verlag, München



Figure 50: *Greek jewelry.*
(Left) Gold spiral bracelet of two snakes whose tails are tied in a Hercules knot that is decorated with a garnet in a bezel setting; from Eretria, on the island of Euboea, 4th–3rd century BC. In the Schmuckmuseum Pforzheim, Germany. (Right) Gold earring with Nike driving a chariot; from the Peloponnese, 4th century BC. In the Museum of Fine Arts, Boston. (Bottom) Gold diadem embellished with blue, green, red, and white enameled flowers; from a tomb at Canossa, 3rd century BC. In the Museo Nazionale di Taranto, Taranto, Italy.

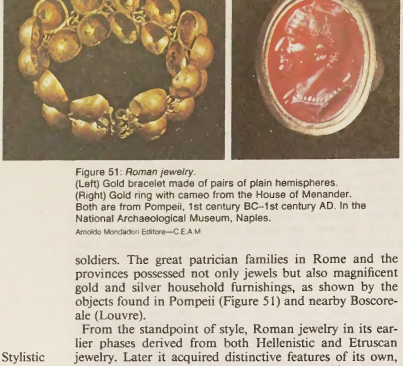


Figure 51: *Roman jewelry.*
(Left) Gold bracelet made of pairs of plain hemispheres.
(Right) Gold ring with cameo from the House of Menander.
Both are from Pompeii, 1st century BC–1st century AD. In the National Archaeological Museum, Naples.
Arnoldo Mondadori Editore—C.E.A.M.

soldiers. The great patrician families in Rome and the provinces possessed not only jewels but also magnificent gold and silver household furnishings, as shown by the objects found in Pompeii (Figure 51) and nearby Boscoreale (Louvre).

From the standpoint of style, Roman jewelry in its earlier phases derived from both Hellenistic and Etruscan jewelry. Later it acquired distinctive features of its own, introducing new decorative themes and attaching greater importance to sheer volume (such as massive rings), in keeping with the rather pompous rhetorical spirit displayed at that point in cultural history.

Stylistic characteristics of Roman jewelry

The motif of a serpent coiled in a double spiral, copied from Hellenistic models, was frequently used for bracelets, rings, arm bands, and earrings. The Romans also used Greek geometric and botanical motifs, palmettos, fleeting dogs, acanthus leaves, spirals, ovoli, and bead sequences. From Etruscan gold jewelry the Romans took the strong plasticity of the bulla, which they transferred to necklace pendants sparely decorated with filigree or combined in completely smooth hemispheres in bracelets, headdresses, and earrings.

In Pompeii and Rome, jewelry began to take on Italian characteristics. New decorative motifs of a magical nature began to appear, such as the half-moon and the wheel with four spokes. In addition, as Roman jewelry freed itself of Hellenistic and Etruscan influences, greater use was made of coloured stones—topazes, emeralds, rubies, sapphires, and pearls. A strong preference was shown for engraved gems, so much so that they were considered collectors' items by wealthy people, including Caesar himself. The stones were set in bezels or supported by pins that passed through them. New techniques that came into use included *opus interassile,* with which a flat or curved metal surface was decorated with tiny pierced motifs, and niello, a method of enameling used primarily to decorate rings and brooches.

Many pendants were used in the earrings: from a ring a series of pieces hung down with square bezels or bands of small bullas alternating with stones, which in turn supported pendants in different shapes. There was an extremely varied production of gold mesh and chains, often containing inserted bezels set with stones or half pearls, while others had ivy or laurel leaves attached to them. Although pendants were not used on necklaces in the beginning, later examples have pendants in the form of embossed medallions. Precious stones, vitreous pastes, and cameos with golden frames also served as pendants for necklaces. Toward the end of the 3rd century AD, necklaces often bore medallions or gold coins with portraits of the emperors.

MIDDLE AGES

Byzantine. Ancient Rome, which had brought its civilization to practically all of the world that was known at that time, began to lose its vitality in the early Christian era; by the end of the 4th century AD, its civilization was in full decline. Although its power was gone, Roman culture was indelibly imprinted on Western civilization. The Roman Empire had embraced Christianity, although in reality it was the papacy that had embraced the Roman Empire. The intention of the Byzantine court (at Constantinople, the new seat of imperial power) to maintain Roman supremacy in the field of the arts was forced to give way to a style more closely related to that of the Middle East. Partly for religious reasons, this style soon developed a new spirit and its own distinctive characteristics. The wave of iconoclasm—the controversy in the 8th and 9th centuries about the depiction of images in religious art—gave the decoration of jewelry, too, a basically ornamental nature, in which the techniques used to the greatest extent were filigree, *opus interassile,* and enameling, as well as the copious application of precious stones and pearls. Very complex decorations and arabesques were obtained with filigree, while enameling was favoured for representations of flowers and birds. Typically Byzantine were the half-moon-shaped earrings that were in wide use up through the 12th century. There are examples with pierced decoration, with filigree basketwork, and with the figures of enameled birds facing each other on a golden half-moon (Figure 52). The court jewels, if credit can be given to the figures shown in the mosaics in the church of San Vitale at Ravenna, must have been of astonishing splendour. Although the mosaics give only a sketchy idea, on the figures of Justinian, Theodora, and their retinue, precious ornaments can be distinguished that were of ceremonial magnificence suited to their rank.

Iconoclasm

For all practical purposes, Constantinople's artistic activities came to an end when it was conquered and looted during the Fourth Crusade in 1204.

By courtesy of the trustees of the British Museum

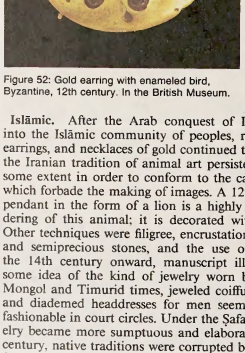


Figure 52: Gold earring with enameled bird, Byzantine, 12th century. In the British Museum.

Islāmic. After the Arab conquest of Iran brought it into the Islāmic community of peoples, rings, pendants, earrings, and necklaces of gold continued to be worn, and the Iranian tradition of animal art persisted, modified to some extent in order to conform to the canons of Islām, which forbade the making of images. A 12th-century gold pendant in the form of a lion is a highly schematic rendering of this animal; it is decorated with granulation. Other techniques were filigree, encrustation with precious and semiprecious stones, and the use of niello. From the 14th century onward, manuscript illustrations give some idea of the kind of jewelry worn by Persians. In Mongol and Timurid times, jeweled coiffures for women and diademed headdresses for men seem to have been fashionable in court circles. Under the Ṣafavid rulers, jewelry became more sumptuous and elaborate. In the 19th century, native traditions were corrupted by European influence, often with an eye toward European consumption. Traditional designs, however, have persisted in Zinjanāb and among the Kurdish mountaineers of northwest Iran. Silver decorated with twisted wire arranged in scrolls is a feature of the former. The Kurdish goldsmiths also work in silver, which they decorate with chased or repoussé

designs, sometimes reminiscent of motifs found on Sāsānian metalwork.

Jewelry worn by men and women in Turkey during the Ottoman period was probably influenced by the fashions current in Iran. Objects of adornment were jeweled turban aigrettes, rings, earrings, necklaces, and armlets. A technique popular in Turkey from the 16th century onward was the encrusting of jade and other hard stones with jewels attached to the surface by delicate floral scrolls in gold. Unfortunately, not many surviving pieces are earlier than the 19th century, when native tradition had been stifled by a taste for Rococo jewelry.

North African designs

In North Africa an independent tradition has been maintained by the Berber and Arab tribes. In design the jewelry of southern Morocco shows curious analogies to Byzantine jewelry—heavy silver plaques decorated with niello or cabochons that serve as diadems or headbands. In other parts of Morocco and in Algeria and Tunisia, popular forms of jewelry are headbands, breast ornaments, brooches, pendants, and a characteristic triangular-shaped shawl pin.

Teutonic. While in the Byzantine area classic forms of expression were being wiped out by the development of a skillful class of artisans who impressed their entirely ornamental taste on jewelry produced solely for decorative purposes, in the rest of Romanized Europe a huge, complex movement of peoples was taking place. Bringing their tradition of polychrome decoration with them, these peoples swarmed over the old declining Greek-Roman artistic civilization. Goths, Vandals, Huns, Franks, and Lombards emigrated, extending their conquests into central, northern, and southern Europe beginning in the 4th century AD, and they remained there until the 9th century. In accordance with an ancient definition, they were called barbarians—that is, not Christians but foreigners. They also were considered barbarians because they were thought to have destroyed the classical art of the Roman world.

Throughout all the provinces of the Roman empire, these Teutonic tribes produced gold ware that shared a common, well-defined style moderated according to the tastes of the particular regions in which they settled. The blend of Teutonic and Iranian, Scythian, Sarmatian, or Celtic styles produced ornaments that bore little resemblance to those of the great classic tradition. Precious ornamentation, which represented the main artistic ambition of these nomadic peoples, was achieved with faience (decoration made of opaque coloured glazes), jewels, and enamels. Dominant also was braiding, which was done with strips of embossing, with bands of stones or enamel set in bezels, and also with filigree.

Teutonic fibulae

There was a highly varied production of fibulae. One of the most impressive for its size (14 inches) is the one in the shape of a bird found in Petroasa, Rom. (National Museum of History, Bucharest, Rom.), whose body is covered with sockets of different sizes and shapes in which stones and enamel were meant to be set (Figure 53). The most widely used type of fibula was the so-called buckler variety, with a fan head, arched bridge, and flat or molded foot, with pierced work in various shapes. Equally common were disk fibulae, either flat or with concentric embossing, while S-shaped fibulae and belt buckles were rarer.

Rigid necklaces, made up of several circles with much decoration, were typical. The most magnificent examples are those from the 6th century from Alleberg and Färjestaden, Sweden (Museum of National Antiquities, Stockholm). A ring with zoomorphic braiding (Poldi Pezzoli Museum, Milan) was found in the same region. This technique was most widely used in the Celtic and northern Germanic regions of Europe, while in the British Isles, to judge from the magnificent jewels in the Sutton Hoo burial-ship treasure (British Museum, London), it was the technique of enameling that reached extremely high levels. In northern Europe and Scandinavia the main goldworking techniques were filigree, embossing, and turning on a lathe.

As time passed, the different products of barbaric goldworking art took on a more definite stylistic identification according to the various races and locations.

Western European. The widespread adoption of Christian burial rites put an end to the custom of burying the dead with all their jewelry. Thus, beginning with the 8th century, almost the only important gold products handed down to modern times were those preserved in abbey and cathedral treasures or by imperial and royal courts; among these gold products are very few pieces of jewelry. As the graphic and plastic arts gradually developed, however, they documented the jewelry in use at the time. According to these sources, little jewelry was worn in the Romanesque period (*c.* 950–*c.* 1150).



Figure 53: *Early medieval jewelry.*
Fibula modeled into the shape of a bird from gold sheet and originally set with stones and cloisonné enamel; Petroasa treasure, 4th century. In the National Museum of History, Bucharest, Rom.
By courtesy of the National Museum of History, Bucharest, Rom.

In the 11th century, monastic workshops for the service of the church began to decline, disappearing one after another to be replaced by secular workshops. Goldworking activities in western Europe gradually freed themselves from the centralizing patronage of the church in order to serve the numerous courts and noble families, and in the 12th century the first goldsmiths' guilds were organized.

Formation of goldsmiths' guilds

One of the most widely used ornaments in medieval Europe was the ring. To it was attributed ever more symbolic and religious value, as well as ever greater importance as a talisman, good omen, and sign of office; and, as always, it served as a seal.

Another widely used ornament was the brooch. Most popular was the medallion type, which might be round, star-shaped, or pentagonal, while the diamond shape was less common. Ring brooches, which were open in the centre, also were popular. They took many forms, including round, pentagonal, and star-, heart-, or wheel-shaped. One outstanding bejeweled and enameled example—the Founder's Jewel bequeathed by William of Wykeham to New College, University of Oxford, in 1404—is in the shape of the letter M (Figure 54, top). The arches formed by the letter resemble Gothic windows, reflecting the importance of architectural elements in all forms of art



Figure 54: *Gothic jewelry.*
(Top) The Founder's Jewel, in the form of a crowned letter "M," the monogram of the Virgin Mary. Gold, emeralds, rubies, pearls, and enamel. English or French, *c.* 1400. In the collection of New College, Oxford. (Bottom) Rhenish belt buckle in gilded silver; from the island of Visby, Sweden, *c.* 1340. In the Historical Museum, Stockholm.
(Bottom) By courtesy of the Historiska Museet, Stockholm. (top) by permission of the Warden and Fellows, New College, Oxford, photograph (top), Thomas-Photos, Oxford

at this time. Standing in the windows are the expertly modeled figures of the Virgin Mary and the Angel of the Annunciation, and the whole is surmounted by a crown.

Flamboyant Gothic style

Another fine example that typifies the plastic decorative repertory of the flamboyant Gothic style is a silver belt buckle from Sweden (Historical Museum, Stockholm). Modeled in high relief on the buckle plate is a gentleman on horseback approaching a lady followed by his servant (Figure 54, bottom). The three-lobed buckle ring is modeled in a complex design that includes a seated person and a man kneeling in front of him (*c.* 1340).

RENAISSANCE TO MODERN

15th and 16th centuries. The "rebirth" of Classicism, which combined all artistic expression in a single orderly, rational approach, found a fertile creative field in gold jewelry. During the Renaissance the jeweler's art reached truly high levels—particularly in Italy in the grand duchy of Tuscany. Eighteen centuries after the great flowering of Hellenistic jewelry, Italian Renaissance jewelry once again achieved an expressive form worthy of comparison with the figurative arts. There was, in fact, no sharp division between the two. Nearly all the most famous artists responsible for the Renaissance artistic revival—Lorenzo Ghiberti, Filippo Brunelleschi, Antonio and Piero Pollaiuolo, and Sandro Botticelli—served apprenticeships in the goldsmiths' workshops, where gentlemen went to order medallions for their hats and where ladies went to have their jewels set.

Because of their elaborate workmanship, which meant that their artistic value was far greater than the intrinsic value of their materials, many pieces of jewelry have been handed down to modern times in public and private collections. Even more extensive evidence, however, is provided by paintings from this period that show the jewelry worn by both men and women. From portraits by Botticelli and Piero di Cosimo, one can see, for example, that as early as the second half of the 15th century the elaborate decoration of women's hair with precious materials had become a real art, in which goldsmiths and craftsmen carefully worked out every line of the often extremely complicated ornamental design that had to harmonize with the movement of braids or unbound hair (Figure 55).

Increase in the use of jewelry during the Renaissance

During the Renaissance there was an enormous increase in the use of jewelry throughout Europe. The courts of England, France, and Spain, the French duchy of Burgundy, and the Italian duchy of Tuscany indulged in extravagant contests, trying to outdo each other in the display of gold, gems, and pearls, a phenomenon that for centuries had not occurred on such a large scale. The nobility and the rich middle class followed this fashion, and even the youngest scions were covered with jewels, as evidenced by the portrait of the Medici princess by Il Bronzino, as well as many others. Francis I of France surrounded himself with famous artists like Benvenuto Cellini and Leonardo da Vinci. In Paris, artists such as Jean Duvet, Étienne Delaune, and the Fleming Abraham de Bruyn were the outstanding creators of designs for jewelry. Hans Holbein the Younger was the individual who was most responsible for the introduction of Renaissance jewelry from the Continent into England, where he found fertile ground, thanks to Henry VIII's great passion for jewels—a passion surpassed only by that of Elizabeth I (Figure 56). Henry possessed more than one magnificent parure, or set of matching jewelry, designed for him by Holbein, as well as several hundred rings.

As Holbein's portrait of Henry VIII suggests, the custom of wearing bejeweled clothing, which had begun gradually in the 14th century, flourished in the Renaissance (Figure 19). Even hat brims were decorated, with designs in pearls as well as with pendants of great value.

In Holbein's portrait there is also a magnificent example of a popular necklace of the period. It consisted of wide gold bands decorated with embossing that formed medallions, in the center of which were mounted large stones. From the necklace hung a pendant. Only rarely were women content to limit themselves to a single necklace, usually wearing a choker-type necklace made of pearls, with or without a pendant, together with a longer second necklace made of gold, with or without the inclusion of

Lauros—Giraudon from Art Resource

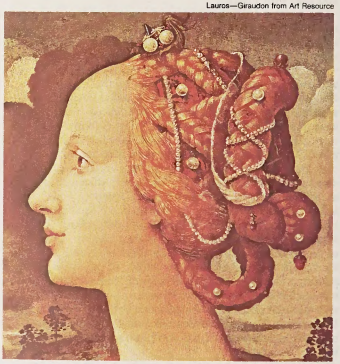


Figure 55: Simonetta Vespucci wearing a pearl headdress with her braids, typical of elaborate Renaissance hair ornamentation. Detail of her portrait, panel painting by Piero di Cosimo; *c.* 1498. In the Condé Museum, Chantilly, Fr.



Figure 56: "Queen Elizabeth of England," adorned in Renaissance fashion with pearl choker and pendant and a series of longer necklaces; portrait in oil by an unknown artist, English, 16th century. In the Pitti Palace, Florence.
Carlo Bevilacqua—SCALA from Art Resource

gems. A third necklace was often hooked to the clothing, on the shoulders, and formed a double loop, being lifted up in the centre and fastened to the bodice with a jeweled pin.

The precious ornament on which the artist-jewelers lavished all their creativity and technical ability was the pendant. At first this consisted of a decorative medallion enclosing a cameo with figures and subjects of classic derivation, such as busts of women and pagan deities. These figures were later enriched with inserts of gold, enamel, and gems, which enhanced the polychrome effect (Figure 57, right). Still later, the figures were freely modeled in brilliant polychromy with a great variety of subjects—animals, Tritons, mermaids, ships, sea monsters, and symbolic figures, sometimes in elaborate tableaux—fashioned in complicated openwork compositions comprising several linked pieces, in which the irregular shape of a large baroque pearl was often used for the body of an animal or a centaur's torso (Figure 57, left).

By courtesy of (left) the Victoria and Albert Museum, London, (right) the Kunsthistorisches Museum, Vienna



Figure 57: *Renaissance pendants.*
(Left) The Canning Jewel, a pendant of gold, enamel, rubies, diamonds, and baroque pearls; German, 16th century. In the Victoria and Albert Museum, London. (Right) Onyx cameo pendant of the goddess Diana wearing a pearl earring enclosed in a gold and enamel frame; Italian, 16th century. In the Kunsthistorisches Museum, Vienna.

Throughout Europe the ring enjoyed wide popularity in an unlimited variety of types, including those with a bezel that could be opened and used as a container for relics, symbols, or—as romantic tradition has it—poison.

17th century. Toward the end of the 16th century, the Renaissance style blended gradually into the manifestations of the Baroque period, which arose at different times in different countries. This gradual change in the style of jewelry was conditioned mainly by two factors. The first was of a technical nature and concerned improvements in the cutting of precious stones, while the second consisted of a great vogue for the cultivation of flowers. Floral and vegetable decoration therefore became the most fashionable theme for jewelry designers, and its popularity spread throughout Europe. The ornamental motifs of knots, ribbons, and Rococo scrolls also saw a considerable development. There was a corresponding decrease in the amount of figurative decoration, which finally completely disappeared. At first these ornamental forms were carried out in openwork gold jewelry, the majority of which was coloured with enamel; later diamonds and other precious stones, whose popularity rose dramatically with the improvement in faceting techniques, became the real protagonists in the composition of jewelry (Figure 58).

Arnoldo Mondadori Editore—C E A M



Figure 58: Baroque earring of pearls, white enamel, and light gold; Spanish, mid-17th century. In the Poldi Pezzoli Museum, Milan.

Period of Louis XIV

During the 17th century the number of pieces of jewelry worn decreased, as did the fashion for male adornment. The last monarch to make heavy use of jewels was Louis XIV, and the word heavy is used here in a literal sense, the great weight consisting mainly of gems with which the monarch covered himself for official ceremonies. He had his own personal jeweler, Gilles Légaré, who was a guest in the Louvre palace. He was not the only sovereign, however, who enjoyed showing off his jewels nor was Versailles the only court in Europe to follow the king's example. Those of London, Madrid, and Munich were

not far behind. The precious ornaments worn by women started on the hat, on the side of which at least one striking aigrette (spray of gems) was fastened. Then came two or three heavy necklaces, each of which might have a pendant, then a belt that followed the pointed shape of the bodice. Other jewels were inserted along the armholes, shoulders, and wrists, and at least four rings were worn on the hands. Often the heavy fabrics used for the clothing were embroidered with gold thread. It was during this period that a spectacular form of jewelry was created in Spain, which in a more subdued form spread throughout Europe: the stomacher brooch, which covered a woman's entire bodice, from neckline to waist. With its heavily bejeweled composition of scrolls, leaves, and pendants on a gold framework that followed the curves of the body, even extending under the armpits, this jewel usually contained no fewer than 50 precious stones of different sizes. A famous example is the one in emeralds from the treasure of the Virgin of Pilar, now displayed in the Victoria and Albert Museum, London (Figure 59).

Use of Brazilian diamonds

18th century. About 1725, Brazilian diamonds in large numbers were imported into Europe, and, during the course of the century, this stone became so popular that imitations were produced. The jewelry of this period seems to have been created to glorify and exploit the cutting of diamonds and other precious stones. The dense forms of Baroque jewelry were replaced by an entirely different conception, in which the design was to appear in gems alone, while the metal setting was concealed to the greatest extent possible. The greater lightness that resulted was increased by the large number of empty spaces in the composition as well as by its lack of symmetry in many cases. Wide choker necklaces with pendants were popular, and the stomacher brooch remained in style but in a lighter, airier form. The jeweled stems of the aigrette were often made so that they could sway back and forth in order to show off the sparkle of the diamonds that covered them. The brooch in the shape of a bouquet of flowers, comprising a variety of gems, became fashionable. As in the 17th century, both men and women wore jeweled buckles on their shoes.

A piece of jewelry that was widely used for daytime wear during this century was the chatelaine, on which, together with the watchcase, goldsmiths lavished some of their most highly refined work (Figure 60). The chatelaine was a pendant made of jointed, embossed gold components of different shapes and sizes, with scenes and designs in elaborate frames. It was fastened by means of a hook to the belt or waistcoat pocket, and from its protruding points hung decorative chains of various lengths, on which men fastened their watches, the keys for winding them, and other accessories. Women used the chatelaine to carry keys, scissors, and other more or less useful objects.

Classical ideals and Josiah Wedgwood

During the last 30 years of the 18th century, the great sensation caused by the archaeological discoveries in Pompeii and Herculaneum caused art forms to turn toward classical ideals of harmony and brought about a decisive change in European tastes and decorative forms. Curved lines no longer appeared in the ornamental repertoire, the new Neoclassical style being characterized by greater simplicity, together with severity of composition. Jewelry forms, too, were influenced by decorative motifs based on Greek and Roman models, and the cameo became fashionable once again.

An English pottery manufacturer, Josiah Wedgwood, made a big contribution to the popularization of the new jewelry forms. An expert technician, he produced reproductions of classic cameos, calling upon sculptors like John Flaxman to work with him on the execution of oval, round, and octagonal plaques with figures done in relief in a white paste on a light blue, green, black, or pink background. These plaques, framed in gold, were used for all sorts of jewelry—medallions, pins, pieces of diadems, belts, bracelets, and rings.

19th century. The Industrial Revolution destroyed forever the ancient role of jewelry as a symbol of social rank. The social evolution created a market for a vast quantity of jewelry at prices the middle class could afford; and

By courtesy of the Victoria and Albert Museum, London



Figure 59: Stomacher brooch with emeralds and enamel flowers on gold; from the treasure of the Virgin of Pilar, Spanish, mid-17th century. In the Victoria and Albert Museum, London.



Figure 60: Chatelaine of gold and enamel; French, late 18th century. In a private collection, Rome.
Arnoldo Mondadori Editore-C.E.A.M.

so jewelry, too, succumbed to the machine. Hundreds of different components for ornaments were produced by machines, an electric gold-plating technique was invented, metal alloys were used in place of gold and silver, and the production of imitation stones increased in both quantity and quality. Despite the growing dominance of the machine, however, the goldsmiths' technical ability remained at a high level.

The jewelry produced in the 19th century is characterized by a stylistic eclecticism that takes its inspiration from all past styles—Gothic, Renaissance, Greek, Etruscan and Roman, Rococo, Naturalistic, Moorish, and Indian, all tinged with the Sir Walter Scott–Lord Byron Romanticism of the period. The futility of transferring forms of artistic expression from an era in which they were the result of organic aesthetic development and of adopting them for objects that reflect only a gesture of romantic admiration is evident in the painting by Jacques-Louis David (Louvre, Paris) immortalizing Napoleon's coronation ceremony in 1804 (Figure 61). The painting provides documentation on the precious ornaments worn by the ladies who were present. In their jewelry, the conventional, rhetorical Empire style appears as a strict, uninspired interpretation of classical motifs, a far cry from the exquisite Neoclassicism of the 18th century.

Artistic commercial jewelry firms

Besides mass production, the 19th century saw the establishment of large artistic commercial firms that produced high-quality jewelry suited to the requirements of the prosperous new bourgeois class. While always satisfying very high standards in regard to technique and materials, these firms tended, from the aesthetic point of view, to reflect the tastes of a bourgeois clientele, which are usually quite traditional.

The oldest of the firms was the one founded by Peter Carl Fabergé in St. Petersburg in 1870, which took over from the firm his father had started in 1842. Fabergé attained great renown at the Universal Exposition in Paris in 1900, where for the first time he put on display all the imperial Easter eggs that he had created, together with a selection of other "luxurious objects." Fabergé used a greater variety of precious and semiprecious stones than almost any other jeweler in history. He had a strong preference for the Louis XVI style but also made numerous objects in the Italian Renaissance, Rococo, and medieval styles, as well as in the so-called old Russian style, which is a mixture of Byzantine and Baroque elements. Decoration with enameling, too, was one of the main specialties of the Fabergé firm.

In Paris in 1898 Alfred Cartier and his son Louis founded a jewelry firm of great refinement. The firm was distinguished for a production characterized by very fine settings, largely of platinum, which were designed so that only the precious stones, always selected from the very purest, were visible (Figure 62, right). At the beginning of the 20th century, Cartier was the most famous jeweler in the world, supplying jewels to the king of Portugal, the duke of Saxe-Coburg-Gotha, the grand dukes and princes of Russia, the Prince of Wales, and other notables.

In the United States in 1851 Charles Lewis Tiffany (father of Louis Comfort Tiffany, one of the most original of the Art Nouveau artist-craftsmen) began producing silverware according to English "sterling" standards in New York City. In 1886 he introduced the Tiffany setting, a special type of fork for the setting of diamonds. Among his clients was President Abraham Lincoln.

Other high-quality jewelry firms founded in the 19th century were Van Cleef & Arpels in Paris (Figure 62, left), Bulgari in Rome, Asprey & Company in London, Black, Starr & Frost in New York City, and Patek Philippe in Geneva.

Art Nouveau

The development of the movement called Art Nouveau at the end of the 19th century represented a reaction against the imitation of ancient styles and the emphasis given, in the creation of jewelry, to precious stones. The material used for Art Nouveau jewelry was prized not for

J.E. Bulloz



Figure 61: Empire style tiaras and garland diadems, necklaces, earrings, and jeweled belts; detail from "The Coronation of Napoleon," oil painting by Jacques-Louis David, 1804. In the Louvre, Paris.

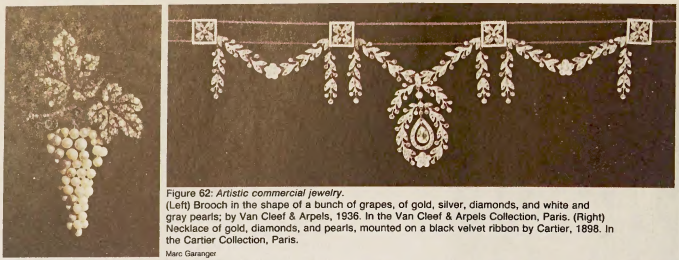


Figure 62: *Artistic commercial jewelry.*
(Left) Brooch in the shape of a bunch of grapes, of gold, silver, diamonds, and white and gray pearls; by Van Cleef & Arpels, 1936. In the Van Cleef & Arpels Collection, Paris. (Right) Necklace of gold, diamonds, and pearls, mounted on a black velvet ribbon by Cartier, 1898. In the Cartier Collection, Paris.
Marc Garanger

its intrinsic value but for its ability to render a design or to carry out chromatic effects. The new jewelry was made from any material that would best express the new symbolic or decorative ideas. Vegetable and animal components, together with the feminine figure, formed the basis for compositions made of flowing lines of rich plastic and chromatic effect and antistructural, dynamic design on a high artistic level.

In Paris, through the works he presented at the Salon du Champ de Mars in 1895, René Lalique (1860–1945) achieved a position of European renown and importance. Lalique personified the Art Nouveau jeweler-artist, his works providing evidence of such highly personal taste that they can be compared to Renaissance jewels. They lean toward a symbolism carried out by the use of milky or watery blue-green colours; of stones such as the opal; of disquieting animals such as the snake, the owl, the octopus, and the bat; and of feminine figures, usually enigmatic, mysterious, and dreamy. Enamel, ivory, vitreous paste, and engraved glass were often used by Lalique to obtain pictorial and plastic effects in his jewels (Figure 63, left).

Unlike Lalique, the jewelers Georges Fouquet (1858–1929) and Henri Vever (1854–1942) expressed themselves through more synthetic geometric forms. The pendant representing a butterfly by Fouquet and the bracelet and ring for the actress Sarah Bernhardt (both in the Périnet Collection, Paris) show a carefully thought-out stylization (Figure 63, right).

The Czechoslovak graphic designer Alphonse Mucha (1860–1939), who worked in Paris, created a number of jewelry designs, transferring his brilliant talent as an illustrator to precious stones and metals.

In the United States the floral style in jewelry found one of its most highly personal interpreters in Louis Comfort Tiffany (1848–1933), one of the greatest of all American designers. In the creation of jewelry he expressed himself at first by transferring to Art Nouveau forms the colourful Oriental and Byzantine style that so fascinated him. Later he adopted Lalique's French Symbolism, on which he set his own stylistic mark. His development of the richly coloured, iridescent Favrile glass created an international sensation.

20th century. The Art Nouveau movement came to an end at the beginning of World War I. The years that followed the war's end seethed with new excitement. In this new phase, the stylistic trends—particularly the nonfigurative—that began to emerge in the most advanced jewelry creations were closely linked to those of painting and sculpture. Cubism, Futurism, the abstractionism of Piet Mondrian and other artists of the de Stijl group, Paul Klee's paintings, and above all the Bauhaus school (which aimed at integrating artistic disciplines with one another and with industrial techniques) provided a basis for the new forms used in avant-garde jewelry.

Compositions were based mainly on the interplay of geometric forms. Like Art Nouveau jewelry, creations of the Art Deco movement (named for the art displayed at the 1925 Paris exposition) used materials suitable for expressing the new stylistic language. Preference was given to the smooth, polished, satined surfaces of precious metals or even of steel. Diamonds and other precious stones were used sparingly, functioning largely as chromatic accents. In the same piece of jewelry, coral could be combined with diamonds, regardless of the great difference in intrinsic value, because their sole purpose was to satisfy the aesthetic requirements of the nonfigurative styles. Art Deco

During this period there were outstanding artist-jewelers such as Raymond Templier, Jean Fouquet, and René Robert in France, H.G. Murphy in England, and Wiwen Nilsson in Sweden.

Later, artists of great international renown devoted some of their creative efforts to the art of jewelry. Some—such as Georges Braque, Jean Cocteau, Max Ernst, Jean Arp, Man Ray, Salvador Dali, Yves Tanguy, and Jean Dubuffet—designed jewelry, while others—including Pablo Picasso,

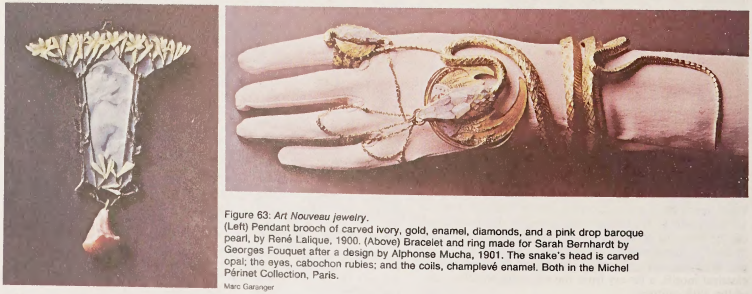


Figure 63: *Art Nouveau jewelry.*
(Left) Pendant brooch of carved ivory, gold, enamel, diamonds, and a pink drop baroque pearl, by René Lalique, 1900. (Above) Bracelet and ring made for Sarah Bernhardt by Georges Fouquet after a design by Alphonse Mucha, 1901. The snake's head is carved opal; the eyes, cabochon rubies; and the coils, champlevé enamel. Both in the Michel Périnet Collection, Paris.
Marc Garanger

Alexander Calder, Alberto Giacometti, Gio Pomodoro—designed and made jewels.

One of the most recent developments in modern mass-produced jewelry is the use of plastic. This material, as well as providing colour, can have mineral fragments or dust embedded in it or can be used in combination with more or less valuable metals, producing pieces of jewelry whose composition may call for considerable effort and which may be of much interest.

NON-WESTERN CULTURES

Jewelry to adorn the costume rather than the body

East Asian. *Chinese.* Much of Chinese jewelry, both of recent and early date, displays the familiar manipulative skill of the Chinese craftsmen; yet the work of the goldsmith or lapidary applied to personal ornament does not represent so distinct a branch of craft as it does in the West and is accorded no special attention by the native connoisseurs and writers on the arts. Most of the jewelry is designed to adorn the costume rather than the person, and much of it has a fulsome and insubstantial quality that is not immediately pleasing to Western eyes. Necklaces, bracelets, and earrings are comparatively rare, headdresses and elaborate hairpins being the more common forms attached to the person (Figure 64). In the traditional costume of recent times, ornate hooks and buckles were used to attach girdles, and women wore strings of beads, often multiple and variously spaced, with decorative plaques and other larger ornaments interspersed. The beads might be attached to the neck, head, or waist, and their purpose was to dignify the whole figure, rather than to display the fine quality of a curiously wrought gem. In any case, the splendour of the stuff of the costume, with richly woven or embroidered ornament, provided the distinctions of rank and wealth, and jewelry was often dispensed with altogether. The long sleeves and high collar of the garment left little of the person exposed for ornament set against the skin, in the manner favoured in the West.

In the time of the Shang dynasty, in the last centuries of the 2nd millennium BC, bone and ivory hairpins with ends carved in the form of birds or abstract figures were a popular adornment. The many finely wrought, small jade plaques of the period, depicting animals in profile, are in many cases clearly intended for sewing to the costume. The earliest evidence of gold ornaments belongs to the time about 400 BC, though these are harness mounts, or weapon parts, rather than jewelry in the usual sense. The latter is better represented by the belt hooks (said to have been adopted from the nomads of inner Asia) that were probably worn by both men and women. They were mostly made of bronze, with fine cast ornaments usually of abstracted dragon and bird heads. These belt hooks were inlaid with gold or silver foil, polished fragments of turquoise, or more rarely with jade or glass; sometimes they were gilded.

Toward the end of the Han dynasty, probably not before the later 2nd century AD, the art of granulation was communicated to China from the Hellenized region of the Black Sea coast. Granulation can be traced in China until about the 10th century AD, its discontinuation in the East curiously coinciding with the loss of the technique in the West. Granulation was combined with filigree; and hairpins, combs, earrings, and costume plaques survive in some quantity, particularly from the richly furnished tombs of the T'ang dynasty (AD 618–907). There are plaques with birds and flowers delineated by soldered wire, inlaid with turquoise, on a ground of fine granulation that appears like a dust of gold.

The employment of the repoussé technique in gold and silver, particularly on the heads of combs, can be attributed to the T'ang period but became more common in the Sung dynasty (AD 960–1279). Meanwhile, hairpins of filigree, with heads shaped as butterflies or flowers, sometimes with pearls or small jade additions, continued the age-old fashion. A scented hairpin takes the place of the scarf or ring of European romance. They were called *pu yao* ("shaking while walking") and were loosely made so as to sway when the wearer moved. Gilded bronze and silver were the principal materials. There are accounts of elaborate headdresses, some no doubt of the kind representing a complete phoenix such as are to be seen on clay tomb statuettes of the T'ang period, but no surviving examples of these can be attributed with certainty to the Sung period. Jade ornaments during this period were still attached to the costume.

Ming dynasty jewelry

Jewelry survives in greater quantity from the Ming dynasty (AD 1368–1644) and gives an impression of greater taste for elaborate figural and floral designs in high repoussé relief and for the effect of semiprecious stones. The latter were prized for their colour rather than their luminosity or rarity. They are never elaborately faceted, being merely ground flat and beveled at the edge for the most part and are set nearly always *en cabochon,* with barely a preliminary polishing, sometimes even retaining the irregularities of the pebble. The stones are invariably semiprecious or even commonplace: amethyst, agate, chalcedony, pink and other quartzes, and, of course, jade. Until modern times, this last has been the most admired of the stones, especially the white variety, which was used for spacers and linking pieces in the silk and beaded hangings of elaborate costumes. The plaques of silver repoussé with flowers and scenes of people were probably used only by men as belt ornaments. Apart from the signet ring, the use of which may not go back beyond Ming times, the male could affect jewelry only in his accoutrement.

© The Board of Trustees of the Victoria and Albert Museum



Figure 64: Ornamental hairpins and earrings of kingfisher feathers and semiprecious stones on silvergilt. Chinese, Ch'ing dynasty, late 19th century. In the Victoria and Albert Museum, London.

Magatama, inrō, and netsuke

Japanese. From as early as 1000 BC until the 6th century AD, Japanese jewelry primarily consisted of comma-shaped objects—not usually more than an inch in length—carved initially of green jade and eventually of glass. Called *magatama,* these beads or pendants were sometimes pierced to be strung in a necklace. The symbolic meaning of the *magatama,* which were often placed in tombs, can only be guessed at. Similar beads also were popular in Korea from the 3rd to the 6th century AD.

In historical times, traditional Japanese costume, male and female, has never allowed the use of ornaments of precious metal or stone, so that nothing in the history of Japanese craft and taste corresponds to the jeweler's work of the West. Hairpins with elaborate heads were increasingly used in the Tokugawa (Edo) period (1603–1868) by women of the geisha and courtesan classes but not by women of other classes. In the same period men were permitted the ostentation of the *inrō,* a small tiered box for tobacco, medicines, confections, and the like, which might be beautifully painted in lacquer and inlaid with mother-of-pearl or precious metal, often in strikingly naturalistic designs. The ivory girdle toggle called netsuke, always delicately and often intriguingly carved, was the only other personal ornament that usage allowed.

Indian. The Indian subcontinent consists of the Republic of India, Pakistan, Bangladesh, and Sri Lanka, but at various times in history its domain has spread to include the neighbouring countries of Nepal, Myanmar, and parts of Afghanistan as well. The area's earliest known urban civilization is called the Indus, or (after an important archaeological site) the Harappān, civilization. It is dated roughly from 2300 to 1750 BC. From this period can be attributed a graceful bronze statue representing a naked dancer. The dancer's hair is braided and decorated, and she wears a necklace with three pendants. Her left arm is fully covered by armlets, and her right arm has an armlet at the elbow and another one near the wrist. This absolutely outstanding specimen provides documentation for the early establishment of the Indian practice of wearing multiple bangle bracelets. Although archaeological evidence of rings, bracelets, and other types of jewelry have been found, no other actual documentation of the way the pieces were worn is available for this period.

Bronze, stone, and ivory sculptures have been discovered dating from the 2nd century BC onward. These include two female figures found in Bhārhut. The statues are lavishly adorned with jewelry: hair ornaments, earrings, necklaces with round and cylindrical beads, chains, belts, coiled ankle bracelets, arm bracelets, and arm rings. Apart from these jewels, the figures wear only a small cloth on their heads. This abundance of jewelry, complemented by little more than veils and scarves, is typical as far as Indian ceramics, painting, and sculpture are concerned. Female figures in Indian art of all periods are almost always depicted wearing huge quantities of jewelry in place of real garments; indeed, the jewels can be thought of as serving as a type of clothing.

The use of jewelry as clothing

The first date for which there is extensive documentation on jewels is the 4th–5th century AD. This information is provided by Buddhist statues and the cycles of wall paintings in the Ajantā caves. Although certainly not the only source for such works, the Ajantā site is one of the most extensive and best preserved. The great variety of types of jewelry represented and the dominance of polychromy indicates the high degree of development attained by the art of jewelry making.

The lavish use of polychrome jewelry was possible because of the ancient practice of pearl diving and because of the wealth and variety of deposits of precious and semiprecious stones to be found in India and the neighbouring countries of Sri Lanka, Thailand, and Myanmar. This situation of plenty, in combination with a favourable climate, helped goldsmiths and jewelers to proliferate and spread, albeit to the detriment of a truly high-class artistry. Although the jewelers were exceptionally skilled craftsmen, they do not seem to have been stylistic innovators. There are no records of particularly gifted artist-jewelers; the only names that have come down through the ages are those of large numbers of patterns.

In the Indus areas and in those under their influence, the setting, polishing, and piercing of precious and semiprecious stones underwent precocious evolution. Stone-cutting, however, was accepted only recently; in the past it was considered preferable not to decrease the size of the stone. In general, there was a preference for a many-hued rich effect that was less a form of artistic self-expression than a display of showy glitter aimed at astonishing the onlooker.

Kinds of Indian jewelry

During the Mughal Empire (1526–1761), rich rajas adorned themselves with jewels—on their turbans, on their ears, around their necks, inserted in their nostrils and even between their teeth. The precious objects worn by women were even more numerous. By this time Indian jewelry had acquired special meanings and nomenclature in connection with a variety of religious beliefs; thus every object had its own specific name, indicating its role and form. For the head alone there were golden wreaths, large brooches, braids made from three bands of gold leaves with a star in the middle set with gems, braids to be placed along the part in the hair, lotus leaves made of gold sheet to be worn at the nape of the neck with bunches of gold flowers next to them, and tiaras in complicated shapes complete with many tinkling pendants. There were similarly large numbers of individually named ornaments for the forehead, the ears, the nose, the neck, the upper part of the arm, the wrist, the fingers, the ankles, and the toes. A variety of forms were used for the earrings, in which pearls, filigree, gems, and coral appeared in floral compositions based on the contrast between the different colours. Some Indian women embedded a jewel in the forehead or pierced the nose in order to wear a jewel in the left nostril. Necklaces were sometimes so long that they came below the navel, and different names were given to those made only of pearls and those of gold. The former also were distinguished according to the number of strings, of which there could be as many as several dozen. Some necklaces were made of a combination of precious stones and pearls, while others were made of amulets in various shapes. A very early type of Hindu amulet called a *nauratan* was made of a gold plaque with nine precious stones fastened above it. A series of *nauratans* could be used to form a necklace. Jeweled belts followed the shape of the body and often had extra pieces that reached up to the neck or down to the bracelets worn around the thigh. Ankle bracelets were often linked by tiny decorative chains running down the instep to the rings on the toes.

Jewelry continues to play an important role in modern Indian dress, but frankly the items produced today do not compare with those of the past. On the contrary, the modern ornaments, though lavishly produced, are of only limited artistic interest.

Southeast Asian. There is a long gold-working tradition among the peoples of Southeast Asia, whose jewelry shows evidence of Tibetan, Chinese, and Indian stylistic influence. The areas in which personal ornamentation with precious objects underwent the greatest development were Myanmar, Cambodia, Laos, and Vietnam. Myanmar jewels are outstanding for the beauty of their designs and for the technical accomplishment of their workmanship. Typical of them is the conical headdress, reflecting the traditional architectural form of the stupa (Buddhist shrine), and the bejeweled, rigid shoulder decorations with a raised line similar to that of pagoda roofs, worn by dancers in addition to arm and ankle bracelets, belts, and brooches made of gold and coloured stones. Although it has its own distinct characteristics, Myanmar jewelry was heavily influenced by Indian styles, especially in regard to a taste for great abundance; thus, each single jewel, rather than standing out, blends into the overall effect.

Myanmar, Cambodia, Laos, and Vietnam

Cambodia, Laos, and Vietnam were subject to greater Chinese influence because of their geographic position. In these territories, too, the principal documentation for the period when precious ornamentation experienced its most flourishing development is to be found in Buddhist sculpture. The outstanding forms of expression in the art of jewelry were thus linked to religious rites, contributing to the glorification of the figures worshiped by the cult.

Scythian. It is to the Scythians, a seminomadic people



Figure 65: Scythian gold bracelet from the Ziwiye treasure, Persia, 7th century BC. In the Guennol Collection, New York City.

By courtesy of the Guennol Collection; photograph, Metropolitan Museum of Art, New York City

from the Eurasian steppes who moved out from southern Russia into the territory between the Don and the Danube and then into Mesopotamia, that we owe a type of gold production, which, on the basis of its themes, is classified today as animal-style. During the early period (5th–4th century BC), this style appeared on shaped, pierced plaques made of gold and silver, which showed running or fighting animals (reindeer, lions, tigers, horses) alone or in pairs facing each other, embossed with powerful plasticity and free interpretation of the forms. The animal-style had a strong influence in western Asia during the 7th century BC. Such ornaments as necklaces, bracelets, pectorals, diadems, and earrings making up the Ziwiye treasure (discovered in Iran near the border between Kurdistan and Azerbaijan) provide evidence of this Asiatic phase of Scythian gold-working art (Figure 65). The ornaments are characterized by highly expressive animal forms. This Central Asian Scythian-Iranian style passed by way of Phoenician trading in the 8th century BC into the Mediterranean and into Western jewelry.

African. Personal decoration in African cultures usually consists of modest though showy material. The works with a relatively high degree of development come from those areas in which the influence of more advanced Mediterranean and Oriental cultures led to activities of some significance in the field of jewelry. Silver was the metal most commonly worked, especially in the northern coastal territories, and the forms used for ornaments were derived mainly from the art of Islām. Decoration that rarely surpassed the level of craftsmanship appears on objects such as bracelets, necklaces, rings, brooches, earrings, and belt buckles, and the techniques were usually limited to embossing, filigree, and the insertion of coins or semi-precious stones that had simply been polished.

Regions such as Ethiopia, the Sudan, and the Bantu territory, partly because of their Egyptian-Nubian and Arabian origins and partly because they were the centres of a flourishing gold trade, developed a gold-working activity of fairly high quality, which was devoted mainly to the production of objects for the courts and for religious ceremonial use. These regions also were devoted to the production of personal ornaments such as embossed plaques, rings, necklaces, and tiaras.

Jewelry of the Ashanti culture of Ghana

The same observations hold true for the Ashanti culture in Ghana, from which there is a large collection of gold jewelry in the British Museum in London. The local chieftain of each Ashanti tribe had a private workshop for gold jewelry in his small court. In the 18th and 19th centuries the most magnificent court was that of the Asantehene (king of the united Ashanti state) in Kumasi, the Ashanti capital on the Gold Coast. A widely used object was the emblem of the "bearer of souls," a decorated disk that, together with other insignia, was borne by the king's pages (Figure 66). On the back of the disk was a little tube through which a gold wire or cord was run. The decoration of these disks consisted of a mixture of separate and varied embossed radial or spiral motifs, derived in an unorthodox manner from classical art. The mysterious presence of these ornamental motifs in Ashanti jewelry can be explained only by the sporadic appearance of European goldsmiths in that area, probably during medieval times. Rigid necklaces also were in use, as were rings, which instead of the bezel had fully sculptured figures of animals.

In the past the sandy dunes of Senegal provided alluvial soil from which the natives obtained much gold. There, as in other parts of Africa, the metal that was not exported was used to make ornaments for the tribal chieftains. These were very elaborate objects with complicated decorative motifs worked in embossing or punched freehand. The objects were characterized by the repetition of the designs used and by protruding hemispheres that were smooth or decorated, according to their size.

In these regions, where the making of jewelry was directly dependent on what was obtained from local deposits, gold was the only material used. Usually the type of decoration, taken from imported models or introduced directly by European goldsmiths, persisted as a repertoire was acquired, with a tendency toward ever greater repetition. In other words, rather than an art form dominated by genuine native expression, this production has no relation to the local culture.

By courtesy of the trustees of the British Museum; photograph, John Webb



Figure 66: Repoussé gold disk depicting the "bearer of souls," Ashanti (Ghana), 18th–19th century. In the British Museum.

American Indian. *Central and South American: pre-Columbian.* The ancient peoples located in the region near the northern Andes (including Peru, Colombia, Ecuador, and Venezuela) achieved a high degree of artistic evolution. Gold mines were abundant in this area, and the goldsmith's art was highly developed. The gold was worked not only by itself but also in alloys with copper, silver, and other metals. The oldest surviving products, attributed to the Chavín culture in Peru, date to as early as 1000 BC. The subsequent Moche culture (*c.* 400 BC–AD 500) and the Nazca culture (both in Peru) also produced gold ornaments of high technical quality.

Andes region

By the time Spanish explorers reached the region in the 16th century, they were astounded at the wealth and magnificence of the then-flourishing Inca empire. Unfortunately, the Spanish melted down most of the gold objects they found. Many examples remain, however—most of them discovered in graves. Study of these materials

has revealed that there were several different centres of production and local styles.

The richest gold and mineral deposits, which are still productive, were those in Colombia. It is not possible to establish definite dates for jewelry from Colombia and Ecuador, but an approximate chronology indicates the San Augustin zone as the oldest, followed by Chitcha. In the latter area, the "Quimbaya treasure" and objects from the upper Cauca River (Calima style) represent jewelry of the greatest importance and magnificence. Other significant centres in Colombia include the Muisca region; Calima, famous for its breastplates, tiaras, and brooches; and Tolima. Although not strictly part of the Andes region, the Coclé region in Panama was strongly influenced by the Quimbaya style. It is particularly known for its striking gold pieces set with precious stones, including emeralds, quartzes, jaspers, opals, agates, and green serpentines.

In the civilizations of the Andes, gold was lavishly used on clothes. About 13,000 pieces of gold were found sewn into a single poncho from Chimú, Peru. On certain occasions the priests wore tunics made entirely of braided gold sheet applied to the cloth. One of the commonest ornaments worn by important personages and warriors was the *nariguera*, a gold ornament that was hooked to the nostrils and might be in the shape of a simple ring, a laminated disk, or an upside-down fan decorated with pierced work. The elite also wore pendants depicting gods or animals.

The most adorned and decorated section of the body was the head. Although gold and other precious metals were components of these ornaments, feathers and other brightly coloured materials were the most important features—the more elaborate the trimmings, the higher the social rank and class of the wearer. Examples of such headdresses can be seen in the great sculptured reliefs found in some ceremonial places.

Rings were rather rare, but there are necklaces with a seashell motif in different shapes arranged one after the other and necklaces with other stylized zoomorphic forms that are all alike. One of the most outstanding of these necklaces is from Chimú (May 21, 1968, Christie sale). It is composed of a row of gold beads to which are attached eight similar figures of a deity in a ritual pose (1100–1200).

Meso-America

Outstanding artistic development during the pre-Columbian era also took place in the region known as Meso-America (including about half of present-day Mexico, all of Guatemala and Belize, and parts of Honduras and El Salvador). When the Spanish reached this area in the early 1500s, they found magnificent monuments, which had been partly invaded and destroyed by woods and brush, but extremely few and scattered people. The reasons that induced the early inhabitants to abandon those places are still unknown, and many potentially illuminating written documents were destroyed by the Spanish. Nevertheless, historical research has determined that the region was inhabitated from about 1500 BC first by the Olmec, then by the Maya, Mixtec, and other groups, and eventually—and until the time of the Spanish conquest—by the Aztec.

Only a few examples of jewelry from this region survive, namely some finely carved jades, which apparently were considered more precious than gold. Works of the goldsmiths' art are rare, although of a high quality. A few examples owned by the Museum of the American Indian in New York City are noteworthy, especially a wonderful Mixtec necklace that proves the high degree of technical skill attained. The necklace is composed of 40 small segments in the form of a tortoise's back, and from each segment hang three drop pendants.

Mixtec graves have yielded outstanding examples of objects such as gold pendants (Figure 67), jewels combined with turquoise mosaic, and quartz ear spools. The few examples that remain from the Aztec period suggest the stylistic influence of the Andes region. Of the decorative animal motifs, the most frequent is the serpent; of the ornamental motifs, the spheroid, disk, and sphere. Probably because the Meso-American area was poor in gold, objects made of this material date from about 1,000 years after those from the Andes (*c.* 14th century BC).

It is thought that ornamental objects in precious materials from the pre-Columbian civilizations, especially the older ones, had some religious function in addition to being used in burial rites. Stylistically, pre-Columbian objects show an unusual amount of charming expressiveness. Symbolic concepts were transferred from stone and pottery to gold through transfigurations that enhanced the plasticity of the forms, displaying at the same time an awareness of structure and of compositional rhythms that forms the main appeal of these objects.

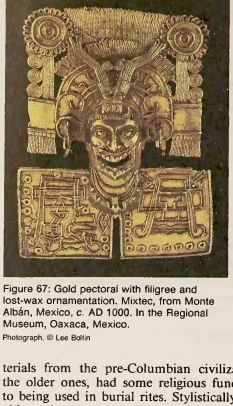


Figure 67: Gold pectoral with filigree and lost-wax ornamentation. Mixtec, from Monte Albán, Mexico, *c.* AD 1000. In the Regional Museum, Oaxaca, Mexico.
Photograph, © Lee Boltin

Religious function of pre-Columbian jewelry

North American. The diverse forms taken by personal ornamentation are related to the type of life led by the numerous ethnic and tribal groups scattered throughout the vast American territory. The most highly developed tribes were those whose social organization permitted them to settle in one place for long periods of time, with the consequent evolution of religious and artistic activities.

On the basis of archaeological finds, North American Indian territory was divided culturally into the following broad areas: the eastern forests, which includes the Great Lakes region and Florida, east of the Mississippi; the Great Plains, including the central part of the continent between the eastern forests and the Rocky Mountains; the Southwest, which corresponds to what are now the states of Arizona, New Mexico, southeastern Utah, and southern Colorado; the northwestern coast, from the bay of Yakutat in Alaska to the mouth of the Columbia River; and California, in the area included between the northwestern coast and the southwestern cultures.

The Great Plains and California produced no jewelry, the former area because its slow artistic evolution involved primarily the decoration of clothing with leather and beadwork, and the latter because its tribes were economically at the preagricultural level and therefore lacking in forms of artistic expression apart from those associated with perishable materials. Judged on the basis of the archaeological data that has come to light so far, the highest artistry was achieved by the southwestern cultures, followed by those of the eastern forests and of the northwest.

Personal ornamentation in all the native cultures of North America shows no connection with the pre-Columbian cultures of Central and South America. One of the most striking differences between the two is that in North America copper was much more frequently used than gold. In some parts of North America this metal may have been used before its use became known in the Western world, and at that distant time it was valued like gold.

As far back as the Archaic period, the practice of decorating shells with carving (Figure 68) or champlevé enamel work was widespread. Feathers and turquoise (used for mosaic) complete the list of precious materials available to the American Indians for personal ornamentation until the arrival of the white man.



Figure 68: Shell etched in the horned toad motif, from the Snaketown excavations, Arizona; Hohokam, AD 900–1150. In the Arizona State Museum, Tucson.

By courtesy of the Arizona State Museum, University of Arizona; photograph, Helga Teiwes

On the whole, in their limited diversity, forms of artistic expression became traditional for particular cultures and were perpetuated by them. Even today, attempts are still being made to keep them alive.

Indians of the Southwest

In the southwestern cultural area the first objects used for personal ornamentation go back to the first half of the 1st millennium AD and consist of bracelets made from a shell carved in the shape of a frog, exquisitely sculptured in miniature; zoomorphic subjects on auricular disks; rings with bird and snake motifs in pierced work; and other shell jewelry covered with turquoise mosaics.

The Pueblo and Navajo tribes, which were part of the southwestern cultural area, made beautiful necklaces and pendants from turquoise mosaics, shells, and coral. The Pueblo Bonito discoveries document this activity from pre-Columbian times. At the beginning of the second half of the 19th century, the Navajo learned to work silver from Mexican craftsmen and developed this skill with great ability, reworking motifs of Spanish American origin in their Indian traditional style.

In the Great Lakes region where the Woodland culture was located, archaeological research has demonstrated the presence of copper ornaments as early as the 5th millennium BC. These consist of necklace beads formed of thin, narrow metal strips and of sheet metal in the shape of fish. The Hopewell finds include bobbin-shaped copper earrings and engraved sheets of silver, dated between 200 BC and AD 400, together with ornaments that were sewn into clothing or inserted in headdresses. From the Mississippian Period there are pieces of embossed copper sheet and breastplates, disks, and plaques made of copper and shell with a wealth of engraved ornamental motifs, such as birds, Sun symbols, isolated heads, human skulls, eagles, rattlesnakes, hands with outspread fingers and an eye designed on the palm, crosses, and figures of warriors.

Beginning in the 17th century, the Seneca, Cayuga, Onondaga, and Iroquois tribes in the New York state region hammered, shaped, and cut European silver coins to be used for jewelry of all kinds. Also worthy of note among the Iroquois are bone combs with handles carved in zoomorphic shapes.

Indians of the northwest coast

In the culture of the Indians on the northwest coast, the influence of Arctic and even of Asiatic peoples can be observed. Persons of very high rank wore a characteristic type of headdress, which was made of wood, in a conical shape with wide brim, surmounted by sculptured human and animal figures. Another type was shaped like a crown or diadem with a rectangular plaque worked in relief placed in the middle of a leather forehead band from which ermine tails and bunches of sea-lion bristles stuck out. The sculpturing on these plaques is highly refined, and the rich shell inlay with which they are decorated makes them look like jewels. The engraving on combs is also outstanding.

The sculptural style peculiar to this culture is characterized by a conventional, formal naturalism that is extremely vigorous and dynamic. Often the same object combines parts that are fully sculptured with parts in low relief, and the depth of the carving may vary greatly.

Objects called copper coins, symbols of maximum power and wealth, were in the form of a shield made of copper sheet in a standardized shape (trapezoidal above and rectangular below). The upper half was taken up by a design such as a head worked in engraving or embossing.

During the 16th century, European conquest and rule of the American continent interrupted development of the arts among the natives, who were forced to live under conditions that were far from favourable to the continuation of traditional artistic activities. (G.Gr.)

BIBLIOGRAPHY

General works. Broad histories of Western dress and clothing accessories, spanning many centuries of development, are FRANÇOIS BOUCHER and YVONNE DESLANDRES, *20,000 Years of Fashion: The History of Costume and Personal Adornment*, new ed. (1987; also published as *A History of Costume in the West*; originally published in French, 1983); MILLIA DAVENPORT, *The Book of Costume*, 2 vol. (1948, reissued in 1 vol., 1976); JAMES LAVER and CHRISTINA PROBERT, *Costume and Fashion: A Concise History*, new ed. (1982); RICHARD CORSON, *Fashions in Makeup: From Ancient to Modern Times* (1972), and *Fashions in Hair: The First Five Thousand Years* (1969, reissued 1984); DOREEN YARWOOD, *European Costume: 4000 Years of Fashion* (1975, reprinted 1982), and *English Costume: From the Second Century B.C. to the Present Day*, 5th ed., rev. (1979); NANCY BRADFIELD, *Historical Costumes of England: From the Eleventh to the Twentieth Century*, 3rd ed., rev. (1970); ALFRED RUBENS, *A History of Jewish Costume*, new and enlarged ed. (1973, reissued 1981); and J. ANDERSON BLACK, MADGE GARLAND, and FRANCES KENNETT, *A History of Fashion*, rev. ed. (1980). More thematically oriented studies include MARGOT HAMILTON HILL and PETER A. BUCKNELL, *The Evolution of Fashion: Pattern and Cut from 1066 to 1930* (1967, reprinted 1987); ELIZABETH EWING, *Fur in Dress* (1981), and *Dress and Undress: A History of Women's Underwear* (1978, reprinted 1989); C. WILLETT CUNNINGTON and PHILLIS CUNNINGTON, *The History of Underclothes*, new rev. ed., by A.D. MANSFIELD and VALERIE MANSFIELD (1981); DIANA DE MARLY, *Working Dress: A History of Occupational Clothing* (1986), and *Fashion for Men: An Illustrated History* (1985); and PENELOPE BYRDE, *The Male Image: Men's Fashion in Britain, 1300–1970* (1979).

Most of the above cited works are well illustrated, but the following two are richly pictorial: DOREEN YARWOOD, *Costume of the Western World: Pictorial Guide and Glossary* (1980); and LUDMILA KYBALOVÁ, OLGA HERBENOVÁ, and MILENA LAMAROVÁ, *The Pictorial Encyclopedia of Fashion* (1968; originally published in German, 1968). Other works suitable for quick reference include DORFEN YARWOOD, *The Encyclopaedia of World Costume* (1978, reissued 1986); C. WILLETT CUNNINGTON, PHILLIS CUNNINGTON, and CHARLES BEARD, *A Dictionary of English Costume* (1968, reissued 1976); MARY PICKEN, *The Fashion Dictionary: Fabric, Sewing, and Apparel as Expressed in the Language of Fashion*, rev. and enlarged ed. (1973); and R. TURNER WILCOX, *The Dictionary of Costume* (1969, reissued 1989).

Ancient dress. Clothes of the early period of civilization are studied in MARY G. HOUSTON, *Ancient Egyptian, Mesopotamian & Persian Costume and Decoration*, 2nd ed. (1954, reprinted 1972), and *Ancient Greek, Roman, and Byzantine Costume & Decoration*, 2nd ed. (1947, reprinted 1966); THOMAS HOPE, *Costume of the Ancients*, new ed., enlarged; 2 vol. (1812, reissued in 1 vol. as *Costumes of the Greeks and Romans*, 1962); LILLIAN M. WILSON, *The Clothing of the Ancient Romans* (1938), and *The Roman Toga* (1924); HANS C. BROHOLM and MARGRETHE HALD, *Costumes of the Bronze Age in Denmark*, trans. from Danish (1940); and ERHARD KLEPPER, *Costume in Antiquity* (1964; originally published in German, 1963).

Medieval dress. Readable and well-illustrated accounts of the developments in Western fashion include STELLA MARY NEWTON, *Fashion in the Age of the Black Prince: A Study of the Years 1340–1365* (1980); MARY G. HOUSTON, *Medieval Costume in England & France: The 13th, and 14th, and 15th Centuries* (1939, reprinted 1965); and JOAN EVANS, *Dress in Mediaeval France* (1952). For non-European medieval dress, see DONALD CORDRY and DOROTHY CORDRY, *Mexican Indian Costumes* (1968, reprinted 1978); SIDNEY M. MEAD, *Traditional Maori Clothing: A Study of Technological and Functional Change* (1969); and TE RANGI HIROA (PETER H. BUCK), *Arts and Crafts of Hawaii* (1957, reissued 1987).

Modern Europe. Histories of fashion include studies of various countries and stylistic trends: see STELLA MARY NEWTON,

The Dress of the Venetians, 1495–1525 (1988); JOHN TELFER DUNBAR, *The Costume of Scotland* (1981); OLGA SROŇKOVÁ, *Fashions Through the Centuries: Renaissance, Baroque, and Rococo* (1959); DIANA DE MARLY, *Louis XIV & Versailles* (1987); ANNE BUCK, *Dress in Eighteenth-Century England* (1979), and *Victorian Costume: And Costume Accessories*, rev. 2nd ed. (1984); MARGARETE BRAUN-RONSDORF, *Mirror of Fashion: A History of European Costume, 1789–1929* (1964; originally published in German, 1963; also published as *The Wheel of Fashion: Costume Since the French Revolution, 1789–1929*); NANCY BRADFIELD, *Costume in Detail: Women's Dress, 1730–1930*, new ed. (1981); NORAH WAUGH, *Corsets and Crinolines* (1954, reissued 1987); MADGE GARLAND, *Fashion* (1962); DIANA DE MARLY, *The History of Haute Couture, 1850–1950* (1980); ELIZABETH EWING, *History of Twentieth Century Fashion*, rev. ed. (1986); and O. E. SCHOEFFLER and WILLIAM GALE, *Esquire's Encyclopedia of 20th Century Men's Fashions* (1973).

For insights into the work of fashion designers, see such surveys as MARTIN BATTERSBY, *Art Deco Fashion: French Designers 1908–1925* (1974, reprinted 1985); IRVING PENN and DIANA VREELAND, *Inventive Paris Clothes, 1909–1939: A Photographic Essay* (1977); and JOHN PEACOCK, *Fashion Sketchbook, 1920–1960* (1977, reissued 1984). See also such biographies and autobiographies as DIANA DE MARLY, *Worth: Father of Haute Couture*, 2nd ed. (1990); PALMER WHITE, *Poiret* (1973); ALFRED ALLAN LEWIS and CONSTANCE WOODWORTH, *Miss Elizabeth Arden* (1972); CHRISTIAN DIOR, *Christian Dior and I* (1957; originally published in French, 1956; also published as *Dior: The Autobiography of Christian Dior*); MARY QUANT, *Quant by Quant* (1966, reissued 1974); and ELSA SCHIAPARELLI, *Shocking Life* (1954).

American dress. Histories of American costume include DIANA DE MARLY, *Dress in North America: The New World, 1492–1800* (1990); ELISABETH MCCLELLAN, *Historic Dress in America, 1607–1800* (1904), and *Historic Dress in America, 1800–1870* (1910), reprinted together as *Historic Dress in America, 1607–1870*, 2 vol. (1990); ALICE EARLE, *Two Centuries of Costume in America, 1620–1820*, 2 vol. (1903, reprinted 1974); ESTELLE ANSLEY WORRELL, *Children's Costume in America, 1607–1910* (1980); and R. TURNER WILCOX, *Five Centuries of American Costume* (1963, reissued 1988).

Oriental dress. Eastern tradition is explored and illustrated in JACQUELINE AYER, *Oriental Costume* (1974); ALAN PRIEST, *Costumes from the Forbidden City* (1945, reissued 1974); SCHUYLER V.R. CAMMANN, *China's Dragon Robes* (1952); A.C. SCOTT, *Chinese Costume in Transition* (1958); SEIROKU NOMA, *Japanese Costume and Textile Arts* (1974; originally published in Japanese, 1965); HELEN BENTON MINNICH, *Japanese Costume and the Makers of Its Elegant Tradition* (1963); G.S. GHURYE, *Indian Costume*, 2nd ed. (1966); and S.N. DAR, *Costumes of India and Pakistan: A Historical and Cultural Study* (1969).

Nature and purpose of dress. Social and psychological aspects of dress and the place of fashion in human culture are the subject of many works representing various disciplines. See JOHN C. FLÜGEL, *The Psychology of Clothes* (1930, reissued 1976); JAMES LAVER, *Modesty in Dress: An Inquiry into the Fundamentals of Fashion* (1969); QUENTIN BELL, *On Human Finery*, 2nd ed., rev. and enlarged (1976); ANNE HOLLANDER, *Seeing Through Clothes* (1978, reissued 1988); ALISON LURIE, *The Language of Clothes* (1981); ELIZABETH WILSON, *Adorned in Dreams: Fashion and Modernity* (1985); and PATRICIA A. CUNNINGHAM and SUSAN VOSO LAB (eds.), *Dress and Popular Culture* (1991). (D.Y./D.J.A. de M.)

Jewelry. Broad surveys include J. ANDERSON BLACK, *A History of Jewels*, rev. ed. (1981; also published as *The Story of Jewelry*); GUIDO GREGORIETTI, *Jewelry Through the Ages* (1969; originally published in Italian, 1969); ERICH STEINGRÄBER, *Antique Jewelry: Its History in Europe from 800 to 1900* (1957); JOAN EVANS, *A History of Jewellery, 1100–1870*, 2nd ed., rev. (1970), and *Magical Jewels of the Middle Ages and the Renaissance, Particularly in England* (1922, reprinted 1976); ANNE WARD *et al.*, *The Ring: From Antiquity to the Twentieth Century* (1981; also published as *Rings Through the Ages*); ERNLE BRADFORD, *Four Centuries of European Jewellery* (1953, reissued 1967); FRITZ FALK, *European Jewellery: From Historism to Modern Style* (1985); and G. MOUREY and A. VALLANCE, *European Art Nouveau Jewelry* (1969).

Materials of the art are studied in C.H.V. SUTHERLAND, *Gold: Its Beauty, Power, and Allure*, 2nd rev. ed. (1969); ROBERT WEBSTER, *Gems, Their Sources, Descriptions, and Identification*, 4th ed., rev. by B.W. ANDERSON (1983); JOHN SINKANKAS, *Gemstones of North America*, 2 vol. (1959–76); and GRAHAM HUGHES, *The Art of Jewelry: A Survey of Craft and Creation* (1972, reissued 1984).

Particular styles and periods are discussed in CYRIL ALDRED, *Jewels of the Pharaohs: Egyptian Jewelry of the Dynastic Period* (1971); REYNOLD HIGGINS, *Greek and Roman Jewellery*, 2nd ed. (1980); JAMILA BRIJ BHUSHAN, *Indian Jewellery, Ornaments, and Decorative Designs*, 2nd rev. ed. (1964); RONALD JESSUP, *Anglo-Saxon Jewellery* (1950, reissued 1974); PRISCILLA E. MULLER, *Jewels in Spain, 1500–1800* (1972); MARGARET FLOWER, *Victorian Jewellery*, new and rev. ed. (1973); JOHN HAYCRAFT, *Finnish Jewellery and Silverware* (1962); and MARY L. DAVIS and GRETA PACK, *Mexican Jewelry* (1963, reprinted 1982).

Books devoted to individual practitioners or describing particular museum collections and exhibitions include A. KENNETH SNOWMAN, *The Art of Carl Fabergé*, 2nd ed. (1962, reissued 1972); HERBERT HOFFMANN and PATRICIA F. DAVIDSON, *Greek Gold: Jewelry from the Age of Alexander* (1965); CHRISTINE ALEXANDER, *Jewelry: The Art of the Goldsmith in Classical Times as Illustrated in the Museum Collection* (1928); EDWARD F. TWINING, *A History of the Crown Jewels of Europe* (1960); MARTIN HOLMES, *The Crown Jewels at the Tower of London*, 4th ed. (1974); HUGH TAIT, *The Waddesdon Bequest: The Legacy of Baron Ferdinand Rothschild to the British Museum* (1981); and HUGH TAIT (ed.), *Jewelry, 7000 Years: An International History and Illustrated Survey from the Collections of the British Museum* (1987; also published as *Seven Thousand Years of Jewellery*). (G.Gr.)

Drugs and Drug Action

Drugs are chemical substances that affect the functioning of living things and the organisms (such as bacteria, fungi, and protozoans) that infect them. Pharmacology, the science of drugs, deals with all aspects of drugs in medicine, including their mechanism of action, physical and chemical properties, metabolism, and therapeutics and prophylaxis.

Until the mid-19th century the approach to drug therapeutics was entirely empirical. This thinking changed when the mechanism of drug action began to be analyzed in physiological terms and when some of the first chemical analyses of naturally occurring drugs were performed. The end of the 19th century signaled the growth of the pharmaceutical industry and the production of the first synthetic drugs. Chemical synthesis has become the most important source of therapeutic drugs, although genetic engineering is being developed as a means of synthesizing proteins.

Drugs produce harmful as well as beneficial effects, and decisions about when and how to use them therapeutically always involve the balancing of benefits and risks. Drugs approved for human use are divided into ethical preparations, available only with a prescription, and over-the-counter drugs, which can be bought freely. The availability of drugs for medical use is regulated by law. Details of all approved substances are published in *The United States Dispensatory* and *The National Formulary*. Similar publications exist in other developed countries.

Drug treatment is the most frequently used type of therapeutic intervention in medicine. Its power and versatility derive from the fact that the body relies extensively on chemical communications systems to achieve integrated function among billions of separate cells. The body is, therefore, highly susceptible to the calculated chemical subversion of parts of this communications network that occurs when drugs are administered.

For coverage of related topics in the *Macropædia* and the *Micropædia*, see the *Propædia*, Part Four, Division II, especially Section 424.

This article is divided into the following sections:

General principles 529
  Mechanism of drug action 529
    Receptors
    Functional macromolecules
    Membrane lipids
    Other types of drug action
  Fate of drugs in the body 531
    Dose–response relationship
    Variability in responses
    Adverse effects
    Absorption, distribution, and elimination
    Time course of drug action
Types of drugs 532
  Autonomic nervous system pharmacology 532
    Mechanism of action
    Sites of action
  Central nervous system pharmacology 535
    Anesthetics
    Analgesics and narcotics
    Drugs affecting mood and behaviour
    Sedative-hypnotic drugs
    Antiepileptic drugs
    Anti-Parkinson drugs
  Cardiovascular system pharmacology 540
    Inotropic agents
    Chronotropic agents
    Antidysrhythmic drugs
    Drugs affecting blood vessels
    Cardiovascular disease
  Drugs affecting blood 542
    Anticoagulant drugs
    Drugs affecting platelets
    Fibrinolytic drugs
  Drugs affecting muscle 544
    Smooth muscle
    Skeletal muscle
  Digestive system pharmacology 545
    Drugs affecting gastrointestinal motility
    Drugs affecting digestive juices
  Reproductive system pharmacology 546
    Female reproductive system
    Male reproductive system
  Kidney pharmacology 547
  Dermatological pharmacology 548
    Topically applied drugs
    Transdermally applied drugs
  Endocrine pharmacology 549
    Anterior pituitary gland
    Posterior pituitary gland
    Adrenal gland
    Thyroid and parathyroid glands
    Pancreas
  Histamine and antihistamines 551
  Chemotherapy 553
    Basic concepts
    Adverse effects
    Antibacterial drugs
    Antifungal drugs
    Antiparasitic drugs
    Antiviral drugs
  Cancer chemotherapy 558
    Alkylating agents
    Antimetabolites
    Antineoplastic antibiotics
    Hormones
    Other agents
    New approaches to cancer therapy
  Immunosuppressants 559
Bibliography 560

General principles

MECHANISM OF DRUG ACTION

With very few exceptions, in order for a drug to affect the function of a cell, an interaction at the molecular level must occur between the drug and some target component of the cell. In most cases the interaction consists of a loose, reversible binding of the drug molecule, although some drugs can form strong chemical bonds with their target sites, resulting in long-lasting effects. Three types of target molecules can be distinguished: (1) receptors; (2) macromolecules that have specific cellular functions, such as enzymes, transport molecules, and nucleic acids; and (3) membrane lipids.

Receptors. Receptors are protein molecules that recognize and respond to the body's own (endogenous) chemical messengers, such as hormones or neurotransmitters. Drug molecules may combine with receptors to initiate a series of physiological and biochemical changes. Receptor-mediated drug effects involve two distinct processes: binding, which is the formation of the drug–receptor complex; and receptor activation, which moderates the effect. Affinity is a term that describes the tendency of a drug to bind to a receptor; efficacy (sometimes called intrinsic activity) describes the ability of the drug–receptor complex to produce a physiological response. Together, the affinity and the efficacy of a drug determine its potency.

Affinity and efficacy

Differences in efficacy determine whether a drug that binds to a receptor is classified as an agonist or as an antagonist. A drug whose efficacy and affinity are suffi-

cient for it to be able to bind to a receptor and affect cell function is said to be an agonist. A drug with the affinity to bind to a receptor but without the efficacy to elicit a response is known as an antagonist because in binding it can block the effect of an agonist.

The binding of a drug to a receptor site requires a precise chemical fit, and a small change in a drug's chemical structure may drastically alter its potency. For example, many drug molecules exist in two forms (called optical isomers), which are identical except that one is a mirror image of the other. One form is usually much more potent than the other, implying that the receptor can distinguish them readily, even though they are chemically almost identical.

The degree of binding can be measured directly by the use of radioactively labeled drugs or inferred indirectly from measurements of the biological effects of agonists and antagonists. Such measurements have shown that the reaction

$$\text{drug} + \text{receptor} \rightleftharpoons \text{drug–receptor complex}$$

generally obeys the law of mass action in its simplest form. Thus, there is a relationship between the concentration of a drug and the amount of drug–receptor complex formed.

The structure–activity relationship describes the connection between chemical structure and biologic effect. Such a relationship explains the efficacies of various drugs and has led to the development of newer drugs with specific mechanisms of action. The contribution of the British pharmacologist Joseph Black to this field led to the development, first, of drugs that selectively block the effects of epinephrine and norepinephrine on the heart (beta blockers, or beta-adrenergic blocking agents) and, second, of drugs that block the effect of histamine on the stomach (H_2-blocking agents), both of which are of major therapeutic importance.

Receptors for many hormones and neurotransmitters have been isolated and biochemically characterized. All of these receptors are proteins, and most are incorporated into the cell membrane in such a way that the binding region faces the exterior of the cell. This allows the endogenous chemicals freer access to the cell. A reconstruction, based on biochemical analysis, on the acetylcholine receptor of muscle is shown in Figure 1. Receptors for steroid hormones (*e.g.*, hydrocortisones and estrogens) differ in being located in the cell nucleus and are accessible only to molecules that can enter the cell across the membrane.

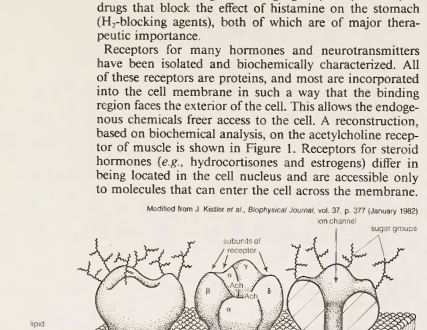


Figure 1: Representation of the acetylcholine receptor, which consists of an aggregate of protein molecules that surrounds a central channel and traverses the lipid layer of the cell membrane.

Receptor-mediated events. Many different kinds of cellular responses may be initiated by receptor activation. For example, nerve cells may be stimulated or inhibited by neurotransmitters acting on surface receptors. In order for the activated receptor to elicit the proper response, certain intermediate processes must take place. Various mechanisms are known to be involved in the processes between receptor activation and the cellular response (also called receptor–effector coupling). Among the most important ones are the following: (1) direct control of ion channels in the cell membrane; (2) regulation of cellular activity by way of intracellular chemical signals, such as cyclic adenosine 3′,5′-monophosphate (cAMP), inositol phosphates, or calcium ions; and (3) regulation of protein synthesis.

Receptor–effector coupling

In type (1) mechanisms, the ion channel is part of the same protein molecule as the receptor, and no biochemical intermediates are involved. Receptor activation briefly opens the transmembrane ion channel, and the resulting flow of sodium and potassium ions across the membrane causes a change in the transmembrane potential of the cell, leading to the initiation or inhibition of electrical impulses. Such mechanisms are common for neurotransmitters that act very rapidly. Examples include the receptors for acetylcholine, which are illustrated in Figure 1, and for other fast excitatory or inhibitory transmitter substances in the nervous system, such as glutamate and glycine.

In type (2) mechanisms, chemical reactions that take place within the cell trigger a series of responses. These mechanisms operate with many hormones and neurotransmitters that act more slowly. The receptor may control calcium influx through the outer cell membrane, thereby altering the concentration of free calcium ions within the cell, or it may control the catalytic activity of one or more membrane-bound enzymes. One of these enzymes is adenylate cyclase, which catalyzes the conversion of adenosine triphosphate (ATP) within the cell to cAMP, which in turn binds to and activates intracellular enzymes that catalyze the attachment of phosphate groups to other functional proteins; these may be involved in a wide variety of intracellular processes, such as muscle contraction, cell division, and membrane permeability to ions. A second receptor-controlled enzyme is phosphodiesterase, which catalyzes the cleavage of a membrane phospholipid, phosphatidylinositol, releasing the intracellular messenger inositol triphosphate. This substance in turn releases calcium from intracellular stores, thus raising the free calcium ion concentration.

Regulation of free calcium ions

Regulation of the concentration of free calcium ions is important because, like cAMP, calcium ions control many cellular functions. An ubiquitous intracellular protein, calmodulin, binds calcium ions, and this complex secondarily controls the activity of various enzymes and other functional proteins.

In type (3) mechanisms, which are peculiar to steroid hormones and related drugs, the steroid–receptor complex acts on specific regions of the genetic material deoxyribonucleic acid (DNA) of the cell nucleus, resulting in the formation of messenger ribonucleic acid (mRNA) that codes for one or more cellular proteins. The effect of the resulting increase in protein synthesis depends on the particular proteins and cell types that are involved. In the uterus, for example, estrogen stimulates the production of structural proteins, leading to growth of the organ. Steroids generally act much more slowly (hours to days) than agents that act by either mechanism (1) or (2).

Desensitization. Most receptor-mediated events show the phenomenon of desensitization, which means that continued or repeated administration of a drug produces a progressively smaller effect. Among the complex mechanisms involved are conversion of the receptors to a refractory (or unresponsive) state in the presence of an agonist, so that activation cannot occur, or the removal of receptors from the cell membrane (down-regulation) after prolonged exposure to an agonist. Desensitization is a reversible process, although it can take hours or days for receptors to recover after down-regulation. The converse process (up-regulation) occurs in some instances when receptor antagonists are administered. These adaptive responses are undoubtedly important when drugs are given over a period of time, and they may account partly for the phenomenon of tolerance (an increase in the dose needed to produce a given effect) that occurs in the therapeutic use of some drugs.

Adaptive responses to drug administration

Functional macromolecules. Many drugs work not by combining with specific receptors but by binding to other proteins, particularly enzymes and transport proteins. For example, physostigmine inhibits the enzyme acetylcholinesterase, which inactivates the physiological transmitter acetylcholine, thereby prolonging and enhancing its

actions; allopurinol inhibits an enzyme that forms uric acid and is used therefore in treating gout. Transport proteins are important in many processes, and they may be targets for drug action. For example, local anesthetics block the conduction of nerve impulses by blocking sodium channels in the nerve membrane, and some antidepressant drugs work by blocking the uptake of norepinephrine or serotonin by nerve terminals. Nucleic acids may also be targets for drug action, as in the case of many anticancer drugs, which prevent cell division by binding to specific regions of DNA.

Membrane lipids. Some drugs produce their effects by chemically nonspecific interaction with membrane lipids. These drugs act nonspecifically in the sense that they affect virtually all cells, irrespective of the presence of any receptor site or other target molecule. Structure–activity relationships within this group show that potency is closely related to lipid solubility and does not depend appreciably on chemical structure. Such chemicals are thought to work by dissolving in the lipid of cell membranes and thus altering their physical properties (*e.g.,* volume and fluidity) in such a way as to affect cellular function. Drugs of this type include volatile anesthetic agents (*e.g.,* halothane) as well as depressants such as ethanol.

Other types of drug action. Certain drugs act without engaging in any direct interaction with the components of the cell. An example is mannitol, an inert polysaccharide that acts purely by its osmotic effect. This drug increases urine production markedly because it interferes osmotically with water reabsorption by the kidney tubule. Another example is magnesium sulfate, which works similarly in the intestine and has a cathartic effect.

FATE OF DRUGS IN THE BODY

Dose–response relationship. The effect produced by a drug varies with the concentration that is present at its site of action and usually approaches a maximum value beyond which a further increase in concentration is no more effective. A useful measure is the median effective dose, ED50, which is defined as the dose producing a response that is 50 percent of the maximum obtainable. ED50 values provide a useful way of comparing the potencies of drugs that produce physiologically similar effects at different concentrations. Sometimes the response is measured in terms of the proportion of individuals in a sample population that show a given all-or-nothing response (*e.g.,* loss of reaction to a painful stimulus or appearance of convulsions) rather than as a continuously graded response; as such, the ED50 represents the dose that causes 50 percent of a sample population to respond. Similar measurements can be used as a rough estimate of drug toxicity, the result being expressed as the median lethal dose (LD50), which is defined as the dose causing mortality in 50 percent of a group of animals.

Median effective dose

When a drug is used therapeutically it is important to understand the margin of safety that exists between the dose needed for the desired effect and the dose that produces unwanted and possibly dangerous side effects. This relationship, termed the therapeutic index, is defined as the ratio LD50:ED50. In general, the narrower this margin, the more likely it is that the drug will produce unwanted effects. The therapeutic index has many limitations, notably the fact that LD50 cannot be measured in humans and when measured in animals is a poor guide to the likelihood of unwanted effects in humans. Nevertheless, the therapeutic index emphasizes the importance of the margin of safety, as distinct from the potency, in determining the usefulness of a drug.

Variability in responses. The response to a given dose of a drug is likely to vary when it is given to different persons or to the same person on different occasions. This is a serious problem, for it can result in a normally effective dose of a drug being ineffective or toxic in other circumstances. Many factors are known to contribute to this variability, with some important ones being age, genetics, absorption, disease states, drug interactions, and drug tolerance.

Adverse effects. No drug is wholly nontoxic or completely safe. Adverse effects can range from minor reactions, such as dizziness or skin reactions, to serious and even fatal effects. Adverse reactions can be divided broadly into effects that result from an exaggeration of the basic action of the drug, which can usually be controlled by reducing the dosage, and effects that are unrelated to the basic action of the drug and occur in only a small proportion of individuals, irrespective of the dose given. Effects of the latter type are known as idiosyncratic effects and include some very severe reactions, such as sudden cardiovascular collapse or irreversible suppression of blood cell production. Many reactions of this type have an allergic basis. Toxic effects of this kind, though rare, are unpredictable and sometimes highly dangerous, and they severely limit the usefulness of many effective drugs. It has been increasingly recognized that drugs can produce other kinds of unwanted effects, such as interference with fetal development (teratogenesis) or long-term genetic damage that may make a person susceptible to the development of cancer.

The sporadic and delayed nature of many adverse drug reactions and the fact that they may not be predictable from animal tests pose serious practical problems. Often such effects are, and indeed can only be, discovered after a drug has been used in humans for some time.

Absorption, distribution, and elimination. In order to produce an effect a drug must reach its target site in adequate concentration. This involves several processes, embraced by the general term pharmacokinetics. In general, the following processes are involved: (1) administration of the drug; (2) absorption from the site of administration into the bloodstream; (3) distribution to other parts of the body, including the target site; (4) metabolic alteration of the drug; (5) excretion of the drug or its metabolites.

An important step in all of these processes is the movement of drug molecules through cellular barriers (*e.g.,* the intestinal wall, the walls of blood vessels, the barrier between the bloodstream and the brain, and the wall of the kidney tubule), which constitute the main restriction to the free dissemination of drug molecules throughout the body. To cross most of these barriers the drug must be able to move through the lipid layer of the cell membrane. Drugs that are highly lipid-soluble do this readily; hence they are rapidly absorbed from the intestine and quickly reach most tissues of the body, including the brain. They readily enter liver cells (one of the main sites of drug metabolism) and are consequently liable to be rapidly metabolized and inactivated. They can also cross the renal tubule easily and thus tend to be reabsorbed into the bloodstream rather than being excreted in the urine.

Lipid solubility

Non-lipid-soluble drugs (*e.g.,* many neuromuscular blocking drugs) behave differently because they cannot easily enter cells. Therefore, they are not absorbed from the intestine and they do not enter the brain. Because they may escape metabolic degradation in the liver, they are excreted unchanged in the urine. Certain of these drugs cross cell membranes, particularly in the liver and kidney, with the help of special transport systems, which can be important factors in determining the rate at which drugs are metabolized and excreted.

Administration. Drugs are given by two general methods: enteral and parenteral administration. Enteral administration involves the esophagus, stomach, and small and large intestines (*i.e.,* the alimentary canal). Methods of administration include oral, sublingual (dissolving the drug under the tongue), and rectal. Parenteral routes, which do not involve the alimentary canal, include intravenous (injection into a vein), subcutaneous (injection under the skin), intramuscular (injection into a muscle), inhalation (infusion through the lungs), and percutaneous (absorption through intact skin).

Enteral and parenteral administration

After oral administration of a drug, absorption into the bloodstream occurs in the stomach and intestine, which usually takes about one to six hours. The rate of absorption depends on factors such as the presence of food in the stomach, the particle size of the drug preparation, and the acidity of intestinal contents. Intravenous administration of a drug can result in effects within a few seconds, making this a useful method for emergency treatment. Serious circulatory and respiratory effects can occur with

this method, however, and great care is necessary. Subcutaneous or intramuscular injection usually produces effects within a few minutes, depending largely on the local blood flow at the site of the injection. Inhalation of volatile or gaseous agents also produces effects in a matter of minutes and is mainly used for anesthetic agents.

Distribution. The bloodstream carries drugs from the site of absorption to the target site and also to sites of metabolism or excretion, such as the liver, kidneys, and, in some cases, the lungs. Many drugs are bound to plasma proteins, and in some cases more than 90 percent of the drug present in the plasma is bound in this way. This bound fraction is inert. Protein binding reduces the overall potency of a drug and provides a reservoir to maintain the level of the active drug in blood plasma. The effects of the drug, therefore, are reduced but prolonged by binding. To pass from the bloodstream to the target site, drug molecules must cross the wall of blood capillaries. This occurs rapidly in most regions of the body. The capillary walls of the brain and spinal cord, however, are relatively impermeable, and in general only drugs that are highly lipid-soluble enter the brain in any appreciable concentration.

Metabolism. In order to alter or stop a drug's biologic activity and prepare it to be eliminated from the body, it must undergo one of many different kinds of chemical transformations. One particularly important site for these actions is the liver. Metabolic reactions in the liver are catalyzed by enzymes located on a system of intracellular membranes known as the endoplasmic reticulum. In most cases the resultant metabolites are less active than the parent drug; however, there are instances where the metabolite is as active as, or even more active than, the parent. In some cases the toxic effects of drugs are produced by metabolites rather than the parent drug.

Role of the liver

Many different kinds of reactions are catalyzed by drug-metabolizing enzymes, including oxidation, reduction, removal of substituent chemical groups, splitting of labile (chemically unstable) bonds, and addition of new substituents. The product is often less lipid-soluble than the parent and is consequently excreted in the urine more rapidly. Many of the causes of variability in drug responses reflect variations in the activity of drug-metabolizing enzymes.

Excretion. The main route of drug excretion is through the kidney; however, volatile and gaseous agents are excreted by the lungs. Small quantities of drugs may pass into sweat, saliva, and human milk, the latter being potentially important in breast-feeding mothers. Although some drugs are excreted mainly unchanged into the urine, most are metabolized first. The first stage in excretion involves passive filtration of plasma through structures in the kidney called glomeruli, through which drug molecules pass freely. The drug thus reaches the renal tubule, where it may be actively or passively reabsorbed, or it may pass through into the urine. Many factors affect the rate of renal excretion of drugs, important ones being binding on plasma proteins (which impedes their passage through the glomerular filter) and urinary acidity, which can affect the rate of passive reabsorption of the drug by altering the state of its ionization.

Time course of drug action. The rise and fall of the concentration of the drug in the blood plasma over time determines the course of action of most drugs. If a drug is given orally, three phases can be distinguished: the absorption phase, leading to a peak in plasma concentration; the redistribution phase, when the plasma concentration falls rapidly as the drug is taken up by various tissues; and the elimination phase, a slower phase of decline as the drug is metabolized or excreted (see Figure 2).

Phases in drug concentration

For therapeutic purposes it is often necessary to maintain the plasma concentration within certain limits over a period of time. If the plasma half-life ($t_{½}$; the time it takes for the plasma concentration to fall to 50 percent of its starting value) is long, doses can be given at relatively long intervals (*e.g.*, once per day), but if the $t_{½}$ is short (less than about 24 hours) more frequent doses are necessary. If a drug with a long half-life is given in intervals over the course of several days, there is a risk that a gradual rise of the concentration of the drug in the plasma will result in toxicity and unwanted side effects.

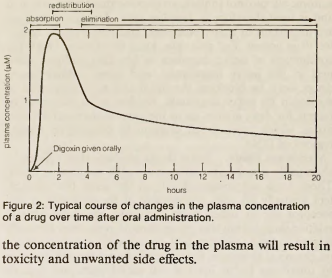


Figure 2: Typical course of changes in the plasma concentration of a drug over time after oral administration.

Types of drugs

AUTONOMIC NERVOUS SYSTEM PHARMACOLOGY

The nervous system of vertebrates comprises two main divisions: the central nervous system, which includes the brain and spinal cord; and the peripheral nervous system, which can be further divided into the somatic nervous system, whose main function is to innervate body structures (*e.g.*, most skeletal muscles) under conscious, voluntary control, and the autonomic nervous system, which is concerned with the involuntary processes of the body's glands, large internal organs, cardiac muscle, and blood vessels. The autonomic nervous system consists of the sympathetic and the parasympathetic systems, which are distinct both functionally and anatomically.

The sympathetic system initiates a series of reactions, called "fight-or-flight" reactions, that prepare the body for activity. The heart rate increases, blood pressure rises, and breathing quickens. The amount of glucose in the blood rises, providing a reservoir of quick energy. The flow of blood to the skin and body organs decreases, allowing more blood to flow to the heart and muscles. The parasympathetic system generally functions in an opposite way, initiating responses associated with rest and energy conservation; its activation causes breathing to slow, salivation to increase, and the body to prepare for digestion. This picture is, however, a considerable oversimplification. The autonomic nervous system as a whole exerts a continuous, local control over the function of many organs (such as the eye, lung, urinary bladder, and genitalia), regardless of whether the body is preparing to react or to rest. The main physiological actions produced by the autonomic nervous system are shown in Table 1.

Table 1: Important Physiological Effects Mediated by the Autonomic Nervous System

organ	response		
	sympathetic activity	receptor type	parasympathetic activity
eye	pupil dilation	α_1	pupil constriction
heart	increase in rate, force, and ectopic beats	β_1	decrease in rate and force
blood vessels	constriction (most)	α_1	dilation (few)
	dilation (some)	β_2	
bronchi	dilation	β_2	constriction
	secretion inhibited		secretion
gut	reduced motility and secretion	α_1, α_2, β_2	increased motility and secretion
bladder	relaxation	β_2	contraction
uterus	relaxation	β_2	contraction
skeletal muscle	glycogen breakdown	β_2	contraction
liver	glycogen breakdown	β_2	– –

Control by the autonomic nervous system

The autonomic nervous system exerts its control through a network of nerve fibres that originate from the cells in the spinal cord. Each of these neurons ends by forming a junction with a second neuron, often called a ganglion cell because in some cases these second neurons are grouped together in swellings called ganglia. The first neuron is

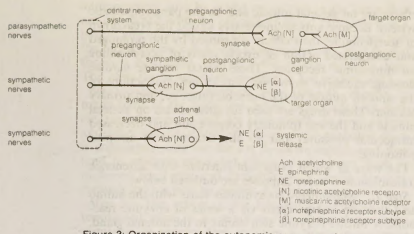


Figure 3: Organization of the autonomic nervous system.

therefore called preganglionic and the second, postganglionic. The junction between the preganglionic and postganglionic neurons is called a synapse. As the electrical nerve impulse reaches the end of the preganglionic neuron, it causes the release of a chemical substance called a neurotransmitter. There is no direct contact between the two neurons. The neurotransmitter diffuses across the gap between them and acts on the postganglionic neuron by initiating in it a further electrical impulse. Postganglionic neurons innervate the target organs and elicit responses in them once again by inducing a neurotransmitter.

Mechanism of action. The discovery of chemical transmitters (neurotransmitters) was an important event in pharmacological history. While a student at Cambridge in 1904, T.R. Elliott found that the effects of stimulating sympathetic nerves closely resembled the effects of injecting chemical substances obtained from the adrenal gland. He suggested that the sympathetic nerves produced their effects by releasing a substance mimicking the action of epinephrine (adrenaline).

The work of Henry Dale, a British physiologist working in London in 1914, suggested that acetylcholine was the neurotransmitter at the synapse between preganglionic and postganglionic sympathetic neurons and also at the ends of postganglionic parasympathetic nerves. He showed that acetylcholine could produce many of the same effects as direct stimulation of parasympathetic nerves. Firm evidence that acetylcholine was in fact the neurotransmitter came in 1921, when the German physiologist Otto Loewi discovered that stimulation of the autonomic nerves to the heart of a frog caused the release of a substance, later identified to be acetylcholine, which slowed the beat of a second heart perfused with fluid from the first. Similar direct evidence of the release of a sympathetic neurotransmitter, later shown to be norepinephrine (noradrenaline), was obtained by Walter Cannon at Harvard also in 1921.

Classification of nerve fibres

In the autonomic nervous system, nerve fibres are classified on the basis of the neurotransmitter released at the synapse. Nerve fibres that release the neurotransmitter acetylcholine are termed cholinergic fibres; nerve fibres that release the neurotransmitter norepinephrine are termed adrenergic fibres. Cholinergic fibres comprise the axons of the preganglionic sympathetic and both the preganglionic and the postganglionic parasympathetic neurons. The axons of the postganglionic sympathetic neurons are generally autonomic adrenergic fibres. The scheme in Figure 3 is complicated by the fact that these neurotransmitters are now known to have a negative feedback effect in inhibiting their own further release. They do this by combining with presynaptic receptors on the nerve terminals as well as with the postsynaptic receptors on the target organs.

Both acetylcholine and norepinephrine act on more than one type of receptor. Dale found that two foreign substances, nicotine and muscarine, could each mimic some, but not all, of the parasympathetic effects of acetylcholine. Nicotine stimulates skeletal muscle and sympathetic ganglia cells. Muscarine, however, stimulates receptor sites located only at the junction between postganglionic parasympathetic neurons and the target organ. Muscarine slows the heart, increases the secretion of body fluids, and prepares the body for digestion. Dale therefore classified the many actions of acetylcholine into nicotinic effects and muscarinic effects. It has subsequently become clear that there are two distinct types of acetylcholine receptors affected by either muscarine or nicotine.

A similar analysis of the sympathetic effects of norepinephrine, epinephrine, and related drugs was carried out by an American pharmacologist, Raymond Ahlquist, who suggested that these agents acted on two principal receptors. A receptor that is activated by the neurotransmitter released by an adrenergic neuron is said to be adrenoceptive or to be an adrenoceptor. Ahlquist termed the two kinds of adrenoceptor alpha (α) and beta (β). This theory was confirmed when Joseph Black developed a new type of drug that was selective for the β-adrenoceptor.

Both α-adrenoceptors and β-adrenoceptors are divided into subclasses: α_1 and α_2; β_1 and β_2 (Table 1). These receptor subtypes were recognized by their responses to specific agonists and antagonists. Once recognized, they provide important leads in developing new drugs with high activity of a certain kind. For example, salbutamol was discovered as a specific β_2-adrenoceptor agonist. It is used to treat asthma and is a great improvement over its predecessor, isoproterenol; because the activity of isoproterenol is not specific, it acts on β_1-adrenoceptors as well as β_2-adrenoceptors, resulting in cardiac effects that are unwanted and sometimes dangerous.

A complex relationship exists between function and receptor type for α-adrenoceptors and β-adrenoceptors. Alpha$_1$-adrenoceptors usually mediate smooth muscle contraction, particularly the constriction of the blood vessels (vasoconstriction) that results from a buildup of calcium ions within the cell. Alpha$_2$-adrenoceptors are located primarily on nerve terminals, where they act to inhibit the release of the neurotransmitter. Beta$_1$-adrenoceptors are found in the heart and increase the force and rate of the heart's action: β_2-adrenoceptors are primarily found in smooth muscle and produce relaxation. Beta-adrenoceptors of both types are involved in the metabolic effects of epinephrine and norepinephrine on liver, fat, and muscle cells, which convert energy stores to freely usable metabolic fuels. The receptor specificity of various agonists and antagonists is summarized in Table 2.

Table 2: Drugs Acting on Cholinergic and Adrenergic Receptors

	agonists	receptor type*	antagonists	receptor type*
cholinergic	acetylcholine	M, N	tubocurarine	N (skel. muscle)
	nicotine	N (mainly ganglia)	hexamethonium	N (ganglia)
	succinylcholine	N (skel. muscle)	trimethaphan	N (ganglia)
			atropine	M
	muscarinic	M	scopolamine	M
	bethanechol	M	homatropine	M
	pilocarpine	M	cyclopentolate	M
adrenergic	norepinephrine	α_1, α_2, β_1	phentolamine	α_1, α_2
	epinephrine	α_1, α_2, β_2	phenoxybenzamine	α_1, α_2
	isoprenaline	β_1, β_2	yohimbine	α_2
	phenylephrine	α_1	prazosin	α_1
			propranolol	β_1, β_2
	salbutamol	β_2	atenolol	β_1
	clonidine	α_2	butoxamine	β_2
			labetalol	α_1, β_1, β_2

* N = nicotinic acetylcholine receptor, M = muscarinic acetylcholine receptor, α_1, α_2, β_1, β_2 = adrenoceptor subtypes.

Other neurotransmitters

It is now known that acetylcholine and norepinephrine are not the only neurotransmitters. There is strong evidence that adenosine triphosphate (ATP), a substance of special importance as a metabolic energy source within cells, also functions as a neurotransmitter in postganglionic autonomic nerves, and it probably mediates some responses (*e.g.*, bladder contraction and vasoconstriction) previously ascribed to acetylcholine and norepinephrine. Dopamine, known to be a metabolic precursor of norepinephrine, is also thought to mediate vasodilator responses in some organs, especially the kidney. A wide variety of peptides, such as substance P, vasoactive intestinal polypeptide, and cholecystokinin, all of which exert powerful effects on target organs, have been detected in autonomic neurons, and it is likely that these also function as neurotransmitters.

Sites of action. Chemical transmission of nerve impulses in the autonomic nervous system involves several steps, some of which are susceptible to interference by drugs: (1) synthesis of the neurotransmitter from simple chemical compounds; (2) storage of the neurotransmitter in a releasable form (generally believed to be in vesicles within nerve terminals); (3) release of the neurotransmitter, which normally occurs when the nerve terminal is invaded by an electrical impulse in the neuron; (4) feedback action of the neurotransmitter on receptors regulating the release of the neurotransmitter; and (5) dissipation of released neurotransmitter by enzymic breakdown or reuptake into nerve terminals. These steps can be used as a basis for describing the effects of drugs on cholinergic and adrenergic transmission.

The ways in which drugs can interfere with cholinergic transmission of nerve impulses are outlined below.

Step 1. Acetylcholine is made from its precursor, choline, which is actively taken up from the blood by cholinergic nerve terminals where an enzyme, choline acetyltransferase, adds an acetyl group to the choline molecule. The uptake of choline can be inhibited by hemicholinium, a substance that blocks all types of cholinergic transmissions. Because of its lack of selectivity, hemicholinium is of no therapeutic value.

Step 2. Acetylcholine storage is not known to be susceptible to control by drugs.

Step 3. Acetylcholine release by nerve impulses can be blocked by botulinum toxin. This very potent toxin is produced in food that is contaminated by the bacteria concerned and is an occasional cause of severe food poisoning. The most serious effect is paralysis of the skeletal muscle, including the respiratory muscle, but it also paralyzes the whole of the autonomic nervous system.

Step 4. There are many drugs that interact with acetylcholine receptors (Table 2). Acetylcholine itself produces extremely short-lived effects because it is rapidly destroyed in the blood. The only acetylcholine-like drug that is employed therapeutically is pilocarpine, a selective muscarinic receptor agonist, which is used in eyedrops to constrict the pupil and to decrease the intraocular pressure that is raised in the disease glaucoma. Nicotine, a constituent of tobacco smoke, stimulates autonomic ganglia but has no therapeutic value.

Antagonists of muscarinic receptors

Antagonists acting on muscarinic receptors include such drugs as atropine and scopolamine. These drugs suppress all of the actions of the parasympathetic system, resulting in drying up of the secretions of the body (*e.g.*, saliva, tears, sweat, bronchial secretions, and gastrointestinal secretions); relaxation of the smooth muscle in the intestine, bronchi, and urinary bladder; an increase in the heart rate; dilation of the pupil; and paralysis of ocular focusing. These drugs are widely used, first to dry up secretions and dilate the bronchi during anesthesia, and second to dilate the pupil during ophthalmological procedures. Scopolamine is also used to treat seasickness, an effect that depends on its ability to depress the activity of the central nervous system.

Nicotinic receptor antagonists are divided into those that act mainly on skeletal muscle and those that act on ganglia cells. The latter group includes hexamethonium and trimethaphan. These drugs cause overall paralysis of the autonomic nervous system because they do not distinguish between sympathetic and parasympathetic ganglia and therefore are not specific in their action. They were the first effective agents to reduce high blood pressure (antihypertensive drugs), but they have a great many troublesome side effects associated with paralysis of the autonomic nervous system (blurred vision, constipation, impotence, inability to urinate). They have been replaced by more selective drugs (see below *Cardiovascular system pharmacology*).

Step 5. Acetylcholine is inactivated by the enzyme acetylcholinesterase, which is located at cholinergic synapses and breaks down the acetylcholine molecule into choline and acetate. Anticholinesterases are drugs that inhibit the activity of this enzyme, thereby greatly enhancing the action of the neurotransmitter. Drugs in this group include neostigmine, Dyflos, and echothiophate. Their main effects are on skeletal muscle and on the parasympathetic system. Because anticholinesterases inhibit the breakdown of acetylcholine, their effects are caused by the prolonged and exaggerated action of excess levels of acetylcholine on muscarinic receptors. These effects include cardiac slowing, excessive secretions, and strong contractions of the smooth muscle of the bronchi, intestine, and urinary bladder. These drugs are used for their effects on skeletal muscle and the eye (pupillary constriction and decreased intraocular pressure) and for the treatment of atropine poisoning.

The ways in which drugs can interfere with adrenergic transmission of nerve impulses are outlined below.

Step 1. Norepinephrine synthesis starts with the amino acid tyrosine and proceeds via a series of enzymic reactions. A further enzymic step occurs in the adrenal gland, where norepinephrine is converted to epinephrine, both of which are released into the bloodstream. Several drugs are known to inhibit one or more of the steps in the synthesis of norepinephrine, but they are not used therapeutically.

Methyldopa affects the synthesis of norepinephrine by causing adrenergic neurons to make and release a false neurotransmitter, methylnorepinephrine, which is less effective than norepinephrine itself. Methyldopa is used in the treatment of high blood pressure and appears to act by affecting parts of the brain that are involved in the regulation of blood pressure and by acting on peripheral adrenergic nerves.

Step 2. The packaging of norepinephrine into vesicles is prevented by reserspine, an alkaloid obtained from the plant *Rauwolfia.* Because of its many side effects, this drug is now largely obsolete.

Step 3. The release of norepinephrine can be evoked or inhibited by the actions of drugs. Drugs that evoke it produce effects resembling those of sympathetic nerve activity and are termed sympathomimetic agents. They include amphetamine and ephedrine, which act indirectly, mainly by expelling norepinephrine from its storage area in nerve terminals. They cause an increase in the heart rate (sometimes leading to dysrhythmias, or irregular heartbeats) and other sympathetic effects. Ephedrine is occasionally used to dilate the bronchi in treating asthma. Amphetamine-like drugs also have strong effects on the brain, causing feelings of excitement and euphoria, as well as a reduction of appetite, which has led to their use in treating obesity. Their effects on the brain have led to their recreational use and to their use as agents to enhance athletic performance. These drugs are liable to cause addiction, and overdosage may have dangerous cardiovascular and mental effects. There is little need for the therapeutic use of the various amphetamines.

Sympathomimetic agents

Step 4. Drugs that act as agonists or antagonists to adrenoceptors are listed in Table 2. Alpha$_1$-adrenoceptor antagonists are important because they block the ability of norepinephrine to constrict the blood vessels (vasoconstriction). Since most blood vessels are subject to the continuous vasoconstrictor influence of sympathetic nerves, blocking these receptors causes a widespread relaxation of the blood vessels (vasodilation). These drugs are sometimes used to treat high blood pressure (hypertension) and cardiac failure (see below *Cardiovascular system pharmacology*).

Beta-adrenoceptor agonists are extremely useful in treating various kinds of cardiovascular diseases, particularly hypertension, dysrhythmias, and angina. The effect is usually achieved by the β_1-adrenoceptor; however, most of the available drugs also block the β_2-adrenoceptor. This gives rise to various unwanted side effects, such as constriction of the bronchial smooth muscle, which can be dangerous to asthmatic patients, and constriction of certain blood vessels, which may cause patients to complain of cold hands and feet. Beta-adrenoceptor antagonists are also useful in controlling muscle tremors and other symptoms of nervousness or anxiety that result from overactivity of the sympathetic system.

Alpha$_2$-adrenoceptor agonists, such as clonidine, are used to treat hypertension and migraine. Clonidine lowers blood pressure by inhibiting the release of norepinephrine from sympathetic nerves, an effect mediated by presynaptic α_2-

adrenoceptors, and by acting on centres in the brain that are concerned with the control of blood pressure. It is a potent and effective drug, but it has the disadvantage that the blood pressure may rise to a dangerously high level if the drug is stopped or even if the patient misses a dose. The basis for its antimigraine effect is not understood.

Beta$_2$-adrenoceptor agonists relax smooth muscle in many parts of the body (see below *Drugs affecting muscle: smooth muscle*) and are used mainly to treat asthma and other allergic disorders. None of the available drugs is completely selective for the β_2-adrenoceptor, and they tend to produce unwanted effects on the heart, such as increased heart rate and disturbances of cardiac rhythm, through their action on cardiac β_1-adrenoceptors.

Step 5. The action of the released norepinephrine is terminated when it is recaptured by sympathetic nerve terminals, a process that involves a selective transport mechanism in the neuronal membrane. Various drugs block this transport system and thus enhance the effects of sympathetic nerve activity; the most important examples are cocaine and certain antidepressant drugs such as imipramine. Overdosage with these drugs results in overactivity of the sympathetic system and the occurrence of cardiac dysrhythmias. The effects of these drugs on brain function, which are of more clinical importance than their peripheral sympathomimetic effects, may be due to this action of inhibiting the uptake of norepinephrine into adrenergic neurons in the brain. (H.P.R.)

CENTRAL NERVOUS SYSTEM PHARMACOLOGY

Anesthetics. Anesthetics are drugs that induce a temporary inability to perceive any sensory stimuli by acting on the brain or peripheral nervous system to suppress responses to sensory stimulation, primarily to touch, pressure, and pain. Such drugs permit the humane application of medical or dental surgical procedures. Anesthesia is the unresponsive state induced by anesthetic drugs. General anesthetics induce anesthesia throughout the body and can be administered either by inhalation or by direct injection into the bloodstream. Local anesthetics provide restricted anesthesia with full retention of consciousness and internal neuronal regulation and are applied to the peripheral sensory nerves innervating a region.

General anesthetics. Inhaled anesthetics act very quickly because of their rapid access into the bloodstream of the lungs and from there directly into the arterial circulation to the brain. The relationship between the amount of general anesthetic administered and the depression of the brain's sensory responsiveness is arbitrarily, but usefully, divided into four stages. Stage I is the loss of consciousness, with modest muscular relaxation, and is suitable for short, minor procedures. Additional anesthetic induces stage II, in which increased excitability and involuntary activity makes surgery impossible; rapid passage through stage II is generally sought by physicians. Full surgical anesthesia is achieved in stage III, which is further subdivided on the basis of the depth and rhythm of spontaneous respiration, pupil reflexes, and spontaneous eye movements. Stage IV anesthesia is indicated by the loss of spontaneous respiration and the imminent collapse of cardiovascular control.

Stages of anesthesia

Not infrequently, general anesthetics are combined with drugs that block neuromuscular impulse transmission. These additional drugs are given to relax muscles in order to make surgical manipulations easier with less suppression of brain activity. Under these conditions, artificial respiration may be required to maintain proper levels of oxygen and carbon dioxide in the blood. The ideal anesthetic agent allows rapid and pleasant induction (the process that brings about anesthesia), close control of the level of anesthesia and rapid reversibility, good muscle relaxation, and few toxic or adverse effects. Some anesthetics have been rejected for therapeutic use because they form explosive mixtures with air, because of their excessive irritant action on the cells that line the major bronchioles of the lung, or because of their adverse effects on the liver or other organ systems.

General anesthesia can be produced by a wide range of chemicals, and most can be administered by inhalation in mixtures with oxygen, permitting the gases to mix with the arterial blood on penetration through the walls of the alveoli within the lung. Rapid induction requires a potent anesthetic with molecular properties that permit efficient transfer between alveolar air spaces and the pulmonary circulation. This rapid route of administration also permits the rapid washout of blood levels of the anesthetic when oxygen alone is administered. Most inhalation anesthetics are excreted by the lungs with little or no metabolism by the body. Safe anesthesia with general anesthetics, such as the intravenously injected barbiturates, requires very short-acting compounds so that the anesthetist can maintain control over the depth of anesthesia while the drug is excreted by the kidney or metabolized by the liver. The mechanism of action of inhalation anesthetics is not well understood.

Nature of most inhalational anesthetics

Except for the naturally occurring gas nitrous oxide, all of the currently used major inhalational anesthetics are hydrocarbons, which are compounds formed of carbon and hydrogen atoms. Each carbon has the potential to bind four hydrogen atoms. The potency of a given series of hydrocarbons depends on the nature of the bonds between the carbons and the degree to which the hydrogen atoms have been replaced with halogens. In the ethers, the carbon atoms are connected through a single oxygen, as in diethyl ether, and again halogen substitution increases potency, as is seen in enflurane, fluroxene, and methoxyflurane. A peculiar, unpredictable, and serious adverse property of halogen anesthetics is their ability to trigger a hypermetabolic reaction in the skeletal muscles of certain susceptible individuals. This potentially fatal response, termed malignant hyperpyrexia, produces a very rapid rise in body temperature, oxygen utilization, and carbon dioxide production.

Rapid, safe, and well-controlled anesthesia can be obtained by the intravenous administration of certain rapidly acting depressants of the central nervous system, such as the barbiturates, the benzodiazepines, or certain synthetic opiates. These systemic anesthetics can be used to produce rapid induction without the discomfort that may accompany induction with the gaseous anesthetics, which are then used to maintain anesthesia. Although opiates such as fentanyl can be used to induce anesthesia, their use is generally limited to very short procedures. Primary advantages are the stability of the cardiovascular system in the presence of this drug and the ability to reverse the effects rapidly with the antagonist drug maloxone. Neither the barbiturates nor the benzodiazepines have any analgesic potency, and barbiturates may, in fact, enhance postsurgical sensitivity to pain. (F.E.Bl.)

Local anesthetics. Local anesthetics produce a loss of sensation in a specific area as a result of their administration into a restricted region, usually by injection. Thus, local anesthetics are useful in minor surgical procedures, such as the extraction of teeth. The first known and generally used local anesthetic was cocaine, an alkaloid extracted from coca leaves obtained from various species of *Erythroxylon*. In the 1880s cocaine was first introduced to the field of ophthalmology for anesthetizing the cornea; later it was used in dental procedures.

The feeling of pain depends upon the transmission of information from a traumatized region to higher centres in the brain. The information is passed along fine nerve (sensory) fibres from the peripheral areas of the body to the spinal cord and then to the brain. If these pain fibres are sectioned, pain sensations from their origins in the periphery are lost. Local anesthetics cause a temporary blocking of conduction along these nerve fibres, producing a temporary loss of pain sensation.

Transmission along nerve fibres

Local anesthetics can block conduction of nerve impulses along all types of nerve fibres, including motor nerve fibres that carry impulses from the brain to the periphery. It is a common experience with normal dosages of an anesthetic, however, that while pain sensation may be lost, motor function is not impaired. For example, use of a local anesthetic in a dental procedure does not prevent movement of the jaw. The selective ability of local anesthetics to block conduction depends on the diameter of the nerve fibres and the length of the fibre that must be affected to block conduction. In general, thinner fibres are

blocked first, and conduction can be blocked when only a short length of fibre is inactivated. Fortunately, the fibres conveying the sensation of dull aching pain are among the thinnest, and the most susceptible to local anesthetics. If large amounts of local anesthetic are used, pain is the first sensation to disappear, followed by sensations of cold, warmth, touch, and deep pressure.

There are many synthetic local anesthetics available, such as procaine, lidocaine, and tetracaine. It is the convention to end the names of local anesthetics with "-caine," after cocaine, which was the first local anesthetic known. In general they are secondary or tertiary amines linked to aromatic groups by an ester or amide linkage. Thus one end of the molecule is hydrophilic ("water loving") while the other end is hydrophobic ("water hating"). The hydrophobic nature of the molecules make it possible for them to penetrate the fatty membrane of the nerve fibres and exert their effects from the inside. When an impulse passes along a nerve, there are transient changes in the properties of the membrane that allow small electrical currents to flow. These currents are carried by ions, especially sodium ions. The influx of these sodium ions through small channels that open briefly in the surface of the nerve membrane during excitation transports the impulse. Local anesthetics block these channels from the inside, preventing the movement of the sodium ions and small electrical currents. The action of a local anesthetic is terminated as the agent is dispersed, metabolized, and excreted by the body. Its dispersal from the injection site depends, in part, on the blood flow through the region. Cocaine, for example, causes blood vessels to constrict, reducing the dispersal rate; other local anesthetics do not have this effect.

Local anesthetics are used to induce limited areas of anesthesia. The limited area is achieved largely by the site and method of administration and partly by the physiochemical properties of the drug molecules. The drug may be injected subcutaneously around sensory nerve endings, enabling minor operations such as tooth extraction to be performed. This is called infiltration anesthesia. Some local anesthetics are applied directly to mucous membranes, such as those of the conjunctiva of the eye or those of the nose, throat, larynx, or urethra. This is called surface or topical anesthesia. A familiar example of topical anesthesia is the use of certain local anesthetics in throat lozenges to relieve the pain of a sore throat. Local anesthetics may be injected near a main nerve trunk in a limb to produce what is called conduction or regional nerve block anesthesia. In this situation, conduction in both motor and sensory fibres is blocked, enabling operations to be carried out on a limb while the patient remains conscious. A special form of regional nerve block may be achieved by injecting a local anesthetic into the spinal canal, either into the space between the two membranes (the durae) that surround the cord (epidural anesthesia) or into the cerebrospinal fluid (spinal or intrathecal anesthesia). In spinal anesthesia, the specific gravity of the local anesthetic solution is appropriately adjusted and the patient is tilted in such a way that the anesthesia is confined to a particular region of the spinal cord. In both epidural and spinal anesthesias, the anesthetic blocks conduction in nerves entering and leaving the cord at the desired level. (A.W.C.)

Infiltration and surface anesthesia

Analgesics and narcotics. Analgesics are drugs that relieve pain selectively without affecting consciousness or sensory perception. This selectivity in relieving pain is the important distinction between an analgesic and an anesthetic. Analgesics are often self-prescribed and abused. Many strong analgesics can produce serious and even fatal depression of the central nervous system. Furthermore, the use of analgesics can involve such problems as addiction, fatal overdose, allergic reaction, and significant gastrointestinal irritation. Analgesics may be classified into two types: the opioids, which act on the brain; and anti-inflammatory drugs, which alleviate pain by reducing local inflammatory responses.

The opioid analgesics were once called narcotic drugs because they induce sleep and cause dependence or addiction. This term is no longer used by physicians or scientists since both dependence and addiction can be induced by many drugs other than opioids. The opioid analgesics can be used for either short-term or long-term relief of severe pain. In contrast, the anti-inflammatory compounds are non-narcotic analgesics and are used for short-term pain relief and for modest pain, such as that of headache, muscle strain, bruises, or arthritis.

Anti-inflammatory analgesics. Several chemically unrelated series of complex organic acids have the ability to relieve mild to moderate pain through actions that reduce inflammation at its source. The prototype of this class of drugs is acetylsalicylic acid, or aspirin. Like aspirin, many of the drugs of this class also reduce fever (antipyretic action), and those that resemble aspirin most closely share what is presumed to be its molecular mechanism of action, namely inhibition of the synthesis of prostaglandins (natural products of inflamed leukocytes, which induce the responses in local tissue that include pain and inflammation). These three properties of anti-inflammatory drugs (*i.e.*, analgesia, anti-inflammation, and antipyretic action) vary among the drugs.

Aspirin

Aspirin, which is now made by total chemical synthesis, is the most widely used mild analgesic, although more potent antipyretic analgesics, such as acetaminophen, are available. Super-aspirins, such as indomethacin, were developed in the 1970s as the mechanisms of critical analgesic actions, namely prostaglandin synthesis inhibition, came to be understood.

Research has shown that small doses of certain prostaglandins can mimic almost all of the signs and symptoms of localized inflammation. Prostaglandins are naturally occurring by-products of arachidonic acid synthesis. They are thought to be released at the site of inflammation when leukocytes are attracted to injured or inflamed areas. All mammalian cells except red blood cells can produce prostaglandins, and when injured, the cells release large amounts of these substances. All aspirin-like analgesics inhibit prostaglandin synthesis and release, and their varying potencies depend on the degree to which they can do so.

As might be expected from their common mechanisms of action, many of the non-narcotic analgesic drugs share similar side effects. Hypersensitivity responses to aspirin-like drugs are thought to be due to an accumulation of prostaglandins after the pathways that break down prostaglandins are blocked. These responses can be fatal when very strong anti-inflammatory compounds are given. Inhibition of prostaglandin synthesis may result in other serious side effects, such as gastric ulcers (which may also be due in part to the irritant activity of large doses of aspirin on the lining of the stomach) and the reduced ability of platelets in the blood to aggregate and form clots. The latter effect, however, has given aspirin an added use as a prophylactic antithrombotic drug to reduce chances of cardiac or cerebral vascular thrombosis. Some of these aspirin-like analgesics also have specific toxic effects: liver damage occasionally occurs after administration of acetaminophen, and renal toxicity and behavioral symptoms are sometimes seen with use of the super-aspirin anti-inflammatory drugs. Aspirin is thought to be a causative agent of Reye's syndrome, a rare and serious degenerative disease of the brain and fatty tissue of the liver that accompanies certain viral infections in children and young adults.

Treatment of mild pain

For treatment of mild pain, two types of analgesics can be used as alternatives to aspirin. The first type comprises the para-aminophenol derivatives such as acetaminophen, which is not a particularly useful anti-inflammatory drug. The term nonsteroidal anti-inflammatory agent is generally reserved for the second type of aspirin-like analgesic, the newer class of highly potent compounds sometimes termed super-aspirins. These drugs all inhibit prostaglandin synthesis. The first of this series, indomethacin, has been largely discarded because of side effects. The most widely used of the super-aspirins are the proprionic acid derivatives, such as ibuprofen, naproxen, and fenoprofen. Although these drugs are widely tolerated and can give significant analgesia in moderate doses, they have also been shown to cause all of the side effects of the other inhibitors of prostaglandin synthesis.

Opioid analgesics. The term opioid has been adopted as a general classification of all of those agents that share chemical structures, sites, and mechanisms of action with the endogenous opioid agonists. Opioid substances encompass all of the natural and synthetic chemical compounds closely related to morphine, whether they act as agonists or antagonists. Although interest in these drugs has always been high because of their value in pain relief and because of problems of abuse and addiction, interest was intensified in the 1970s and 1980s by discoveries about the naturally occurring morphine-like substances, the endogenous opioid neuropeptides.

Opium is the powder from the dried juice of the poppy *Papaver somniferum*. When taken orally, opium produces sleep and induces a state of peaceful well-being. Its use dates back at least to Babylonian civilization. In the early 19th century opium extract was found to contain more than 20 distinct complex organic bases, termed alkaloids, of which morphine, codeine, and papaverine are the most important. These pure alkaloids replaced crude opium extracts in therapeutics.

In the 1950s several new morphine-like drugs were suggested. Despite the increase in the number of compounds available for pain relief, however, little was understood of their sites and mechanisms of action. The first real breakthrough came from the discovery, by J.W. Hughes and H.W. Kosterlitz, of two potent naturally occurring analgesic pentapeptides (peptides containing five linked amino acids) in extracts of pig brain. They called these compounds enkephalins, and since then at least six more have been found. Larger peptides, called endorphins, have been isolated; and these contain sequences of amino acids that can be split off as enkephalins. There are at least three types of receptors on brain neurons that are activated by the enkephalins. Morphine and its congeners are thought to exert their effects by activating one or more of these receptors. This is still a rapidly advancing field of research. It is thought that the common final pathway of receptor activation is the movement of calcium ions through the plasma membrane.

Enkephalins

Opioid drugs are useful in the treatment of general postoperative pain, severe pain, and other specific conditions. The use of opiates to relieve the pain associated with kidney stones or gallstones presumably depends on their ability to affect opiate receptors in these tissues and to inhibit contractility. By a similar mechanism, opiates are also able to relieve the abdominal distress and fluid loss of diarrhea. Central receptors appear to account for the ability of morphine and analogues to suppress coughing, an effect that requires lower doses than those needed for analgesia. Low doses of opiates given subcutaneously are also specifically advocated for the relief of the respiratory distress that accompanies acute cardiac insufficiency complicated by the buildup of fluid in the lungs, even though the mechanisms of this effect are unknown and despite the fact that opiates are respiratory depressants.

Several commonly used natural or synthetic derivatives of morphine are used in drug therapeutics. Codeine, a naturally occurring opium alkaloid, is also made synthetically and provides a useful adjuvant analgesia as an oral preparation, especially when used in combination with aspirin. Meperidine, also known as Demerol, was one of the earlier synthetic analogues of morphine that was originally thought to be able to provide significant short-lasting analgesia and little or no addiction because of its shortened duration of action. This belief proved false and the drug is widely abused. Methadone, a synthetic opioid analgesic, has long-lasting (six to eight hours) analgesic effects when taken orally and the unique ability to antagonize the euphoria-producing effects of heroin; it is therefore used to moderate the effects of withdrawal from opiate addiction. Among the opioid antagonist drugs, naloxone and its longer lasting orally active version, naltrexone, are used primarily to reverse morphine overdoses and to reverse the chemical stupor of a wider variety of causes, including alcohol intoxication and anesthesia. In opiate overdoses, where the signs of pinpoint pupils, depressed respiration, and unconsciousness help in the diagnosis, these drugs provide an almost miraculous recovery within minutes of injection. They can, however, also precipitate severe withdrawal reactions if the subject had been addicted previously.

Methadone

The effectiveness of a given dose of an opiate drug declines with its repeated administration in the presence of intense pain. This loss in effectiveness is termed tolerance. Evidence suggests that tolerance is not due to alterations in the brain's responses to drugs. Animals exhibiting tolerance to morphine after repeated injections in a familiar environment show little or no tolerance when given the same doses and tested for pain sensitivity in new environments. Thus, a learned aspect of tolerance seems almost certain. The cellular and molecular mechanisms underlying this loss of responsiveness are not clear, however. Physical dependence and addiction in a person using intravenous administration closely follow the dynamics of drug tolerance; increasing doses are required to produce the psychological effects, while tolerance protects the brain against the respiratory depressant actions of the drug. In the tolerant individual, intense adverse reactions can be precipitated by administration of an opiate antagonist, thus revealing the dynamic internal equilibrium that previously appeared to neutralize the response of the brain to the opiates. The signs of the withdrawal response (yawning, tearing, perspiration, dilation of the pupils, nasal discharge, anxiety, tremors, elevation of blood pressure, abdominal cramps, and hyperthemia) can be viewed as signs of an activated sympathetic nervous system and to some extent an extreme, but nonspecific, arousal response. This may also suggest that tolerance in the face of increasing opiate doses is more properly viewed as a balancing mechanism by this sympathetic activation.

Drugs affecting mood and behaviour. Behaviour and emotions are higher functional properties of the brain that depend on the network of neurons and chemical neurotransmitters that exist throughout the body; however, the means by which neurons achieve changes in behaviour and in mood remains unknown. Nevertheless, certain neurotransmitters, such as the monoamines, norepinephrine, dopamine, epinephrine, serotonin, and acetylcholine, appear to be closely linked to these aspects of brain function. Drugs that influence the operation of these neurotransmitter systems can profoundly influence and alter the behaviour of patients with psychiatric problems.

Classification

Drugs that affect mood and behaviour can be classified as follows: antianxiety agents, antidepressants, antipsychotics, antimanics, stimulants, antiappetitives, and antiemetics. Such drugs should be reserved for severe disruptions of normal emotional well-being and should not be used to relieve the boredom, tension, or sadness that may be properly regarded as a normal part of life. Sedative-hypnotics, which depress conscious awareness and induce sleep, are described in a separate section (see below *Sedative-hypnotic drugs*).

Antianxiety drugs. Anxiety is a state of pervasive apprehension that may be triggered by specific environmental or personal factors. Anxiety states are generally combined with emotions such as fear, anger, or depression. A person suffering from anxiety may complain of physical symptoms such as palpitations, nausea, dizziness, headaches, and chest pains, as well as sleeplessness and fatigue. When such apprehension is severe and incapacitating, the person is said to suffer from anxiety neurosis, which may require treatment by psychotherapy. Many of the drugs used in the treatment of anxiety are for the most part safe and well tolerated and physicians often prescribe them either as an alternative to psychotherapy in severe cases, or as an aid to coping with different situations in mild cases.

After World War II Swiss pharmacologists discovered muscle relaxant properties in a compound under investigation as an antibiotic. Modification of that compound led to the tranquilizing drug meprobamate. Another discovery showed that the benzodiazepines, which are complex ringed compounds, had even greater relaxing properties. Hundreds of analogues of the basic benzodiazepine ring were subsequently synthesized. The most widely prescribed compounds, chlordiazepoxide and diazepam, are now giving way to shorter acting compounds that are less likely to produce sedation. Different formulations of the basic

benzodiazepine structure in higher dosages are used as muscle relaxants, antiepileptics, and hypnotics (see below *Sedative-hypnotic drugs* and *Antiepileptic drugs*).

The brain exhibits highly specific, high-affinity binding sites that can selectively recognize, or bind, the benzodiazepine compounds. The cellular and subcellular locations of these sites are near ion channels in the membrane that can admit chloride ions into the cell and also near sites where a neurotransmitter, gamma-aminobutyric acid (GABA), acts. Benzodiazepine agonists in general enhance the effects of GABA. In 1985 scientists in the United States showed that brain extracts contain an endogenous inhibitor of benzodiazepine binding. Assessment of its behavioral effects on the brain suggests that this natural compound may cause rather than suppress anxiety and decrease rather than increase GABA transmission.

Acute treatment with benzodiazepines generally begins with doses taken before bedtime to facilitate sleep. Because the need for the drugs depends on the patient's response to psychotherapy and his ability to reshape the events that lead to the anxiety, more or less tolerance may develop to the sedation. There are side effects with the use of benzodiazepines. Because of the alterations in the effectiveness of inhibitory transmitter actions of GABA, which are profound in the cerebellum and cerebral cortex, the patient may also exhibit confusion and loss of motor coordination. Other drugs, especially alcohol, taken with benzodiazepines can interfere with coordination, and use of these drugs during pregnancy may increase chances of fetal malformations.

Side effects of benzodiazepines

Antidepressants. Clinical depression is characterized by a sad or hopeless mood, a loss of interest in one's usual activities, reduced energy, change of appetite, disturbed sleep patterns, and often contemplation of suicide. An individual must experience these symptoms for at least two weeks in order to be diagnosed with clinical depression. The disorder also must be distinguished from grief felt in reaction to the death of a loved one or some other unfortunate circumstance.

In 1957 imipramine emerged as the first therapeutically useful antidepressant. An accidental discovery led to the finding that the drug iproniazid caused some patients to become extremely euphoric and hyperactive by inhibiting monoamine oxidase, a brain enzyme that normally breaks down norepinephrine and other neurotransmitters. Drugs that were better at blocking the activity of this enzyme were even more effective in evoking euphoria. Shortly thereafter, the monoamine oxidase inhibitors (MAOIs), as they were later called, were introduced for the treatment of depression. Another class of antidepressants called tricyclics, named for their basic three-carbon ring structure, were discovered around the same time as the MAOIs. Tricyclics inhibit the active reuptake of the neurotransmitters norepinephrine, serotonin, and dopamine in the brain. Inhibition of reuptake allows the neurotransmitters to remain in contact longer with their postsynaptic receptors. This mechanism seems to support the hypothesis that depression is caused by a chemical imbalance in the levels of neurotransmitters.

The most common antidepressants used today are selective serotonin reuptake inhibitors (SSRIs). By inhibiting only the reabsorption of serotonin, SSRIs have fewer, less serious side effects than the other classes of antidepressants, which interfere with several neurotransmitter systems and may result in increased sensitivity of the sympathetic nervous system and cardiovascular irregularities. Introduced in the late 1980s, SSRIs include fluoxetine (Prozac), paroxetine (Paxil), and sertraline (Zoloft). Other antidepressants such as buproprion (Wellbutrin) that are chemically unrelated to the other antidepressants also were introduced. Antidepressants are also prescribed for the treatment of obsessive-compulsive disorder, panic disorder, and bulimia nervosa.

Side effects vary among the types of antidepressants and may include sleepiness, tremors, anxiety, loss of sexual desire, and nausea. Although some improvement in symptoms may occur in the first week or two, antidepressant medications typically need to be taken for at least three weeks before the full therapeutic effect of the drug occurs. Most physicians recommend that patients continue to take antidepressants for at least six months to prevent a recurrence of symptoms.

Antimanics. Mania is a severe form of emotional disturbance in which the patient is progressively and inappropriately euphoric and simultaneously hyperactive in speech and locomotor behaviour. This is often accompanied by significant insomnia, excessive talking, extreme confidence, and increased appetite. As the episode builds the patient exhibits racing thoughts, extreme agitation, and incoherence, frequently replaced with delusions, hallucinations, and paranoid fears. Ultimately the patient may become hostile and violent, and may finally collapse. In some patients, periods of depression and mania alternate, giving rise to the form of affective psychosis known as bipolar depression, or manic-depressive disorder. The most effective medications for this form of emotional disorder are the simple salts lithium chloride or lithium carbonate. Although some serious side effects can occur with large doses of lithium, the ability to monitor blood levels and keep the doses within modest ranges (approximately one milliequivalent [mEq] per litre) makes it an effective remedy for manic episodes and it can also stabilize the mood swings of the manic-depressive patient.

Lithium salts

In the 1940s lithium salts were used briefly as a possible sodium substitute for subjects with cardiac failure and hypertension until the salts' serious effects on cardiac rhythms became apparent. The use of lithium in the treatment of mania emerged from a series of ambiguous and ill-conceived observations. Lithium was given to manic patients who showed the calming effects now recognized as valid therapeutic responses. The treatment was initially regarded skeptically, and its use was long delayed in the United States. Lithium is now an accepted and preferred form of long-term treatment for both manic and manic-depressive patients.

Given the simplicity of the salt, the fact that therapeutic action resides in lithium itself and not the anion that accompanies it, and that there is no direct metabolism of the lithium, scientists have been unable to determine how the antimanic effects arise. The use of lithium involves a latent period of 10–14 days before the mania subsides, after which, if blood lithium levels are kept constant (within the one mEq plasma concentration), there is generally a good stabilization of the patient's moods. The positively charged lithium ions enter into a biophysical equilibrium with calcium and magnesium, and this replacement action, with slow accumulation in the cells of the brain, is regarded as critical for the stabilization of emotion.

If patients take an overdose, or if their normal salt and water metabolism becomes unbalanced by intervening infections that cause anorexia or fluid loss, then signs of lithium overdose may become apparent. Such signs include loss of coordination, drowsiness, weakness, slurred speech, and blurred vision, as well as more serious chaotic cardiac rhythm and brain-wave activity with seizures. Because lithium is generally excreted along with sodium in the urine, rehydration and supportive therapy are all that are required. Prolonged treatments with lithium, however, can in fact damage the body's ability to respond properly to the antidiuretic hormone, which stimulates the reabsorption of water, thus causing the emergence of diabetes insipidus. Lithium can also interfere with the response of the thyroid gland to the thyroxin-stimulating hormone produced in the pituitary gland. Both the renal and thyroid effects of lithium have been attributed to the blockade of a cell surface response that requires the target cell to synthesize the intracellular mediator cyclic adenosine monophosphate.

Antipsychotic drugs. The severe form of mental illness known as schizophrenia is usually a chronic, often lifelong, inability to think logically and act appropriately. Effective treatments for some forms of schizophrenia have revolutionized thinking about the disease and have prompted investigations into its possible genetic origins and biopathological causes. Of the three most common patterns of schizophrenic psychosis, treatments with antipsychotics are the most useful in the case of young adults who show an acute onset of paranoid fears and halluci-

Treatment of schizophrenia

nations and in middle-aged adults whose childhood and adolescence were marked by poor interpersonal relationships and disturbed thinking.

Each of the three major series of drugs that have been used successfully in the treatment of schizophrenia has a colourful origin. The history of reserpine can be traced to the Indian shrub called *Rauwolfia serpentina* for its snakelike appearance, which historically had been used to treat snake bites, insomnia, high blood pressure, and insanity. Reserpine, the principle alkaloid of the plant, was first isolated in the 1950s and used in the treatment of hypertension; it was later given to disturbed schizophrenics, in whom the drug was found to act as a behavioral depressant. In fact, the depression of patients given the drug for hypertension was a major side effect. The basic mechanisms of action of reserpine in producing depression are attributed to its ability to deplete the brain's stores of the monoamines.

The second major class of antipsychotic drugs, the phenothiazines, arose from modifications of the dye methylene blue, which was under investigation as an antagonist of histamine. Attempts to modify this series to increase their activity in the central nervous system and reduce the need for surgical anesthetics ultimately led to the first effective drug of this class, chlorpromazine. Its ability to stabilize behaviour and to improve lucidity as well as to reduce hallucinatory behaviour was recognized throughout the world within a few years of its introduction. The use of chlorpromazine changed the role of the mental hospital and resulted in the large-scale, perhaps excessive, discharge of medicated schizophrenics into the outside world.

The third class of antipsychotics, the butyrophenones, emerged when a small Belgian drug company embarked on an ill-conceived plan to develop analogues of Demerol (a synthetic narcotic analgesic) through cheap chemical substitutions. Experiments gave rise to a compound that caused chlorpromazine-like sedation but had a completely different structure. This led to the compound haloperidol, a more powerful antipsychotic with relatively fewer side effects.

Drugs related to chlorpromazine and to haloperidol share several actions in humans and in animals. These drugs are sedatives, which means that they can suppress the vomiting reflex and nausea (for this reason they are used in the early stages of pregnancy), and they can improve many of the symptoms of schizophrenia in a significant number of patients. Although its mechanisms of action are not well established, they are thought to involve a deficiency of the monoamine transmitter dopamine. Schizophrenia may be brought about either by an excessive release of dopamine or, more likely, by increased sensitivity to dopamine. Thus the effectiveness of the antipsychotic drugs chlorpromazine and haloperidol, as well as chemically related drugs, may be attributed to their ability to antagonize, or inhibit, the actions of dopamine.

Side effects of chlorpromazine and haloperidol

The major acute side effects of chlorpromazine and haloperidol are oversedation and a malaise that makes the drugs poorly received by the patient and makes compliance with chronic self-medication difficult. On prolonged treatment of middle-aged and even young adults, antipsychotic drugs can evoke serious movement disorders that in part resemble Parkinson's disease, a degenerative condition of the nerves. First to appear are tremors and rigidity, followed by more complex movement disorders commonly associated with involuntary twitching movements on the arms, lips, and tongue, called tardive dyskinesia. The latter disorder also has been described in nonpsychotic subjects who were sedated and in untreated schizophrenic patients. One explanation for this side effect is that these movement disorders may be a part of the natural course of the psychotic disease but can be greatly exacerbated by treatment with the dopamine antagonist drugs. In addition to these often irreversible side effects, there can also be adverse liver, cardiac, and bone marrow toxicity. Because of these problems with long-term treatment and because most schizophrenic patients are helped minimally if at all, the search for the other possible causes of the psychosis and of other forms of antipsychotic therapy has continued.

Sedative-hypnotic drugs. Drugs that reduce tension and calm anxiety at low doses (see above *Antianxiety drugs*) and that produce drowsiness and facilitate the onset of sleep at higher doses are called sedative-hypnotics. Because this state of sleep is one from which a patient can normally be aroused, its production was once attributed to "hypnotic" actions, but the sleep that is induced is actually quite natural. Still higher doses of some sedative-hypnotics can produce deep unconsciousness sufficient to make them useful as general anesthetics.

The dose levels at which calm, sleep, or anesthesia are induced depend on the drug classes and their mechanisms of action. Since similar effects can be obtained with other drugs, such as analgesic opiates or antianxiety benzodiazepines, the principal characteristic of primary sedative-hypnotics is their selective ability to induce these actions without affecting mood or sensitivity to pain.

Alcoholic beverages and alcoholic extracts of opium were traditionally used as sedative-hypnotics, but the first substance introduced specifically as a sedative and as a hypnotic was a liquid solution of bromide salts. In 1869 chloral hydrate became the first synthetic organic molecule to be employed specifically for its sedative-hypnotic effect, and it was followed by several others, notably paraldehyde. (Chloral hydrate was used notoriously as "knock-out" drops.) Barbiturates, with their more complex organic ring structure, were introduced in the early 1900s, and hundreds of barbiturate analogues were then synthesized with varying potencies and durations of action. Potent analogues of barbiturates have been used to induce surgical anesthesia and to reduce voluntary inhibition during psychiatric interviews (for which they have sometimes been dubbed "truth serums"). Most of the barbiturates were discontinued after the development in the 1950s of the benzodiazepines, many of which exhibit the ideal properties of a short-acting, intense facilitator of natural sleep, with a reduced risk of adverse effects.

Depressant effects of alcohol

The means by which alcohol depresses brain function and produces sleep is not clear; however, it is certain that intoxicating doses of alcohol greatly incapacitate the processing of information, reduce reaction times, and depress skilled locomotor behaviours. No single neurotransmitter system has been definitively shown to be the locus of these behavioral effects, and multiple mechanisms may be involved. Barbiturates and benzodiazepines have similar but not identical actions at the behavioral and cellular level to those of alcohol. The benzodiazepines act on the inhibitory sites at which gamma-aminobutyric acid (GABA) is the neurotransmitter.

Effects of alcohol and barbiturates

In certain persons low doses of alcohol, barbiturates, and some benzodiazepines produce transiently enhanced mood or euphoria, along with antianxiety effects. These behavioral effects can lead to abuse of these substances and to dependence upon them with prolonged use. High doses can depress critical centres in the brain stem for the regulation of cardiovascular and respiratory function.

When sedatives are taken frequently as sleeping tablets, tolerance and a reduction in effectiveness occurs. Despite popular beliefs to the contrary, alcoholic beverages in particular are only of modest benefit to induce sleep. On frequent exposure to alcohol the nervous system adapts to the drug, and this results in early morning awakening. Barbiturates can be selected to provide both early onset of sleep and a prolongation of sleep. Analysis of electroencephalographic patterns during barbiturate-induced sleep, however, shows that there is more disruption of sleep. There have been reports that some benzodiazepines used as sleep inducers produce less disruption of the sleep phases, a property that makes them especially useful for persons with sleep disturbances.

Antiepileptic drugs. Epilepsy is a general term for a group of central nervous system disorders characterized by transient but repeated episodes of abnormal electroencephalographic activity that correlate with abnormal motor behaviour (convulsions) and, less commonly, with sensory, autonomic, or psychological manifestations. Although some forms of epilepsy may be caused by high fevers, especially in infants, and while some forms of epilepsy in adults can be traced to previous brain injury, with resulting scars, or to brain tumours, the causes of

Treatment of epilepsy

most forms are unknown. As a result of the uncertain origin, the treatment of epilepsy is directed toward reducing the frequency of seizure. Moreover, many of the drugs that have been found to be useful have led to trial-and-error testing of analogues of barbiturate and benzodiazepine sedatives. An accurate diagnosis of the form of epilepsy is critical to the drug most likely to be effective.

The mechanism of action of most antiepileptic drugs is unknown, except for the barbiturates and the benzodiazepines, which act by enhancing the effectiveness of the inhibitory neurotransmitter substance GABA. Other drugs were discovered by testing their ability to prevent seizures in experimental animals after electrical stimulation of the brain or after the administration of convulsant drugs such as strychnine, pentylenetetrazol, or metrazol. On the other hand, the hydantoins, such as phenytoin, were discovered as a result of persistent testing of a series of drugs. Phenytoin is effective in the long-term treatment of many varieties of epilepsy, especially in combination with more sedative barbiturates.

The tricyclic antidepressant drug carbamazepine, used in the treatment of trigeminal neuralgia, was later found to have value in the treatment of epileptic disorders. It is also effective against some chemically induced seizures. The effectiveness of the drug has been attributed to its ability to enhance monoaminergic transmission.

Because most epileptic conditions are long-lasting and of unknown origin, their treatment is confined to drugs. As might be expected, side effects after prolonged use are common. Phenytoin, for example, may be directly toxic to neurons of the cerebellum, and this may actually exacerbate the seizures. In addition, this drug can cause gingival hyperplasia and hirsutism, which may lead patients to abandon it.

Anti-Parkinson drugs. Parkinson's disease, or paralysis agitans, is a severe progressive degenerative disease of the nerves characterized by tremor of the distal limbs that disappears when movement is initiated. Later in the course of the disease the muscles become rigid, and the initiation of movements and their termination, once started, become so difficult as to be incapacitating. In the final stages the patient is unable to maintain an erect posture, speak, write, or focus the eyes. Although the loss of pigmented neurons of the brain region called the substantia nigra had been a pathological finding in postmortem specimens since the early 20th century, a pathophysiological explanation of the disorder was not found until 1960. These neurons use the substance dopamine as their neurotransmitter, and they project onto the basal ganglia, a centre for the coordination of movement. Patients with Parkinson's disease were found to have basal ganglia greatly deficient in dopamine.

Debilitating course of Parkinson's disease

Recognition that this chemical deficiency of a specific neurotransmitter was a central feature of the disease led to a new therapy based on the use of the amino acid L-3,4-dihydroxyphenylalanine (L-dopa), the precursor of dopamine. When given orally in large daily doses, some L-dopa is able to escape metabolism in the bloodstream and enter the brain, where surviving dopamine neurons convert it to dopamine. Although the degenerative changes are progressive, L-dopa may also be converted to dopamine by a specific enzyme. To increase the delivery of this dopamine precursor to the brain, L-dopa therapy is supplemented with carbidopa, an analogue of L-dopa that inhibits this enzyme in the intestine and in the general circulation but that is unable to penetrate into the brain. As a result, carbidopa increases the effectiveness of L-dopa. Overdosage with L-dopa can cause schizophrenia-like episodes, presumably due to the excess formation of dopamine. Some patients are also helped by bromocriptine (a dopamine-like agonist so modified as to be able to gain access to the brain). The use of L-dopa to treat Parkinson's disease, however, is not the radical cure that it was once thought to be but only a measure that modifies to some extent the degenerative changes of the disease. (F.E.Bl.)

CARDIOVASCULAR SYSTEM PHARMACOLOGY

Drugs that affect the function of the heart and blood vessels are among the most widely used in medicine. Although these drugs may exert their primary effect either on the blood vessels or on the heart itself, the cardiovascular system functions as an integral unit. Thus, drugs that affect blood vessels are often useful in treating conditions in which the primary disorder lies in the heart itself, or vice versa. Examples of disorders in which such drugs may be useful include hypertension (high blood pressure), angina pectoris (pain resulting from inadequate blood flow through the coronary vessels to the muscular wall of the heart), heart failure (inadequacy of the output of the heart in relation to the needs of the rest of the body), and arrhythmias (disturbances of cardiac rhythm).

Drug effects on the heart

Drugs affect the function of the heart in three main ways. They can affect the force of contraction of the heart muscle (inotropic effects); they can affect the frequency of the heartbeat, or heart rate (chronotropic effects); or they can affect the regularity of the heart beat (rhythmic effects).

Drugs affect blood vessels by altering the state of contraction of the smooth muscle in the vessel wall, altering its calibre, or diameter, thereby regulating the volume of blood flow. Such drugs are classified as vasoconstrictors if they cause the smooth muscle lining to contract, and vasodilators if they cause it to relax. Drugs may act directly on the smooth muscle cells, or they may act indirectly, for example by altering the activity of nerves of the autonomic nervous system that regulate vasoconstriction or vasodilation (see above *Autonomic nervous system pharmacology*). Another type of indirect mechanism is the action of vasodilator substances that work by releasing a smooth muscle relaxant substance from the cells lining the interior of the vessel. Some drugs mainly affect arteries, which control the resistance to blood flow in the vascular system, an important determinant of the arterial blood pressure; others mainly affect the veins, which control the pressure of blood flowing back to the heart, and hence the cardiac output (*i.e.*, the volume of blood pumped out by the heart per minute).

Inotropic agents. Inotropic agents are drugs that influence the force of contraction of cardiac muscle, thereby tending to affect the cardiac output. Drugs have a positive inotropic effect if they increase the force of contraction of the heart. The most important group of inotropic agents is the cardiac glycosides, substances that occur in the leaves of the foxglove (*Digitalis purpurea*) and other plants.

Although they have been used for many purposes throughout the centuries the effectiveness of cardiac glycosides in heart disease was established in 1785 by an English physician, William Withering, who successfully used an extract of foxglove leaves to treat heart failure. Many closely related glycosides with similar pharmacological actions are found in various plants, but they differ in ease of absorption from the gastrointestinal tract and in duration of action. The two compounds most often used therapeutically are digoxin and digitoxin.

Effects of cardiac glycosides

The most useful effect of cardiac glycosides is their ability to increase the force of contraction of cardiac muscle. They have, however, several additional effects, most of which are disadvantageous. These include a tendency to block conduction of the electrical impulse that causes contraction as it passes from the atria to the ventricles of the heart (heart block). Cardiac glycosides also have a tendency to produce an abnormal cardiac rhythm by causing electrical impulses to be generated at points in the heart other than the normal pacemaker region, the cells that rhythmically maintain the heartbeat. These irregular impulses result in ectopic heartbeats that are out of sequence with the normal cardiac rhythm. Occasional ectopic beats are harmless, but if this process continues to a complete disorganization of the cardiac rhythm (ventricular fibrillation), the pumping action of the heart is stopped, causing death within minutes unless resuscitation is carried out. Because the margin of safety between the therapeutic and the toxic doses of glycosides is relatively narrow, they must be used carefully.

Cardiac glycosides are believed to increase the force of cardiac muscle contraction by binding to and inhibiting the action of a membrane enzyme that extrudes sodium ions from the cell interior. Inhibiting the free flow of sodium ions from the interior of the cell across the membrane

to the exterior of the cell causes the intracellular sodium concentration to rise. The interior of the cell then becomes depolarized, or electrically less negative than normal with respect to the exterior of the cell. Because the cell is able to exchange sodium ions within the cell for calcium ions outside it, there is a secondary rise in intracellular calcium. This subsequently increases the force of contraction, since intracellular calcium ions are responsible for initiating the shortening of muscle cells.

The disturbances of rhythm that may be caused by cardiac glycosides result partly from the depolarization and partly from the increase in intracellular calcium. Because these rhythm disturbances are caused by the same underlying mechanism that causes the beneficial effect, there is no likelihood of finding a cardiac glycoside with a significantly better margin of safety. Apart from their cardiac actions, these glycosides tend to cause nausea and loss of appetite. Because digoxin and digitoxin have long plasma half-lives (two and seven days, respectively), they are liable to accumulate in the body. Treatment with either of these drugs must involve careful monitoring to avoid the adverse effects that may result from their slow buildup in the body.

The second type of inotropic agent that increases the force of cardiac muscle contraction includes epinephrine and norepinephrine. In addition to affecting the force of contraction, however, they also increase the heart rate. This, and the fact that they are quickly metabolized by the body and act only for a few minutes, means that they are not useful inotropic agents.

Caffeine-related drugs

The third type of inotropic agent that acts as a cardiac stimulant is the caffeine-related series of drugs represented by theophylline. Its action, like that of epinephrine, depends on an increase in the intracellular concentration of cyclic adenosine 3′,5′-monophosphate, which indirectly increases the influx of calcium ions into the cells, thereby increasing the force of contraction of cardiac muscle.

Chronotropic agents. The heart rate is controlled by the opposing actions of sympathetic and parasympathetic nerves and by the action of epinephrine released from the adrenal gland. Norepinephrine, released by sympathetic nerves in the heart, and epinephrine, released by the adrenal gland, increase the heart rate, while acetylcholine, released from parasympathetic nerves, decreases it. A competitive antagonist that acts to inhibit the stimulating action of norepinephrine on the heart is propranolol, which slows the heart and is often used to treat anginal attacks and disturbances of cardiac rhythm. Atropine blocks acetylcholine receptors and is used during anesthesia to prevent excessive cardiac slowing.

Antidysrhythmic drugs. Many types of heart disease lead to disturbances of the cardiac rhythm, a common example being the occurrence of ventricular dysrhythmias following heart attacks. Dysrhythmias are undesirable because they compromise the pumping action of the heart and because they can worsen suddenly and lead to cardiac arrest. The regularity of the heartbeat depends on the activity of the pacemaker area, located in the sinoatrial node of the heart, which generates electrical impulses at a frequency of 70 per minute. The rate is regulated by the opposing influences of the sympathetic and parasympathetic nerves. The impulse from the pacemaker spreads in an orderly sequence to the rest of the heart, resulting in a contraction of the atrial chambers, forcing the blood in these chambers into the ventricles. This is followed about 0.3 second later by contraction of the ventricles, the main pumping chambers, which forces the blood into the arteries.

Reentrant rhythms

The cardiac rhythm can be disturbed in several ways. (1) The conduction pathway may be disorganized, resulting in a reentrant rhythm in which an impulse circulates continuously in a local area of the heart (often a damaged region), causing irregular reexcitation of the rest of the heart at an abnormally high rate. If this happens in the atria it is called atrial fibrillation and the ventricular beat continues, though sometimes with an irregular rhythm. If this irregular heart rate occurs in the ventricles it is called ventricular fibrillation and the pumping action of the heart ceases. (2) Ectopic, or abnormally placed, pacemakers may appear in regions of the heart other than the sinoatrial node, and these can drive the heart at an abnormally high rate. (3) Various forms of heart block can occur in which the impulse fails to continue in the heart at some point because of local damage. This results in slowing or complete cessation of the heartbeat.

Drugs are useful in treating all of these types of dysrhythmia. Reentrant rhythm and ectopic pacemakers cause abnormally high heart rates (tachycardia), and they require treatment with drugs that slow the heart and reduce the electrical excitability of the muscle cells. Reentrant rhythms can be eliminated by increasing the refractory period of the cells, which is the interval following transmission of an electrical impulse during which the cell cannot be reexcited by another impulse. Increasing the refractory period has the effect of reducing the frequency at which impulses can be transmitted.

Quinidine and procainamide (mainly used for arterial dysrhythmias) and lidocaines and phenytoin (mainly used for ventricular dysrhythmias) exert their antidysrhythmic effects by reducing electrical excitability. Quinidine and procainamide have the disadvantage that they reduce the force of contraction of the heart and tend to lower blood pressure. They are also liable to cause side effects such as nausea and skin rashes. Lidocaine, which is also used as a local anesthetic, has a very short duration of action and must be given intravenously; its main use is in the prevention of ventricular dysrhythmias following acute occlusion of a coronary artery.

An important factor tending to exacerbate ectopic pacemakers is the release of norepinephrine from sympathetic nerves. Norepinephrine acts on β-adrenoceptors in the heart to increase its rate, which strongly increases the tendency for ectopic pacemakers to develop. Beta-adrenoceptor blocking drugs (*e.g.*, propranolol) are widely used to control these types of dysrhythmia because they slow the actions of the heart. They also tend to reduce the force of contraction of the heart, which can be a disadvantage, and they produce various other unwanted effects.

Calcium antagonists

In the mid-1970s, the calcium antagonists, another type of antidysrhythmic drug, were introduced. Verapamil and diltiazem are important examples of this class of drugs. They reduce the influx of calcium ions through the cell membrane, which normally occurs when the cell is depolarized. This movement of calcium ions across the membrane appears to be important as a factor in the genesis of reentrant rhythms and ectopic heartbeats. Inhibiting the influx of calcium ions is effective in controlling many types of dysrhythmia. Since calcium entry is essential for initiating the contraction of heart muscle cells, calcium antagonists tend to impair muscle contractility. Since calcium entry is also important in the contraction of blood vessel smooth muscle, calcium antagonists cause vasodilation and tend to lower arterial blood pressure.

All of the antidysrhythmic drugs discussed so far impair the conduction of the impulse for contraction from atria to ventricles and therefore can cause heart block. Antidysrhythmic drugs should be used carefully to avoid the various hazards and side effects that they may produce. Heart block causes a pathological slowing of the heart and is not usually treated with drugs, although β-adrenoceptor agonists such as isoproterenol are sometimes used in emergencies. An artificial electrical pacemaker device is usually fitted to provide effective long-term control.

Drugs affecting blood vessels. Many different drugs and endogenous substances cause constriction or dilation of blood vessels. Drugs that have a direct relaxant effect on vascular smooth muscle include the organic nitrates (*e.g.*, nitroglycerin tablets, which are mainly used to treat angina) and calcium antagonists (*e.g.*, nifidipine). Most blood vessels are controlled by the sympathetic nervous system, and they constrict in response to norepinephrine released from sympathetic nerves. Thus, drugs that affect the sympathetic system (see above *Autonomic nervous system pharmacology*) cause constriction or dilation of blood vessels. The parasympathetic nervous system is much less important in controlling blood vessels.

Apart from the actions of the sympathetic nervous system, several other physiological mechanisms regulate vascular

Renin-angiotensin system

smooth muscle. Of particular pharmacological importance are the renin-angiotensin system and locally acting vasodilator substances, such as histamine, bradykinin, and prostaglandins.

Renin is an enzyme that is released into the bloodstream by the kidney when the blood pressure falls. It acts on a plasma protein to produce a peptide, angiotensin I, which consists of a chain of 10 amino acids. This in turn is acted on by angiotensin converting enzyme (ACE) to produce an eight-amino acid peptide, angiotensin II (a potent vasoconstrictor), which raises the blood pressure. Inhibitors of ACE are used in treating high blood pressure.

Various substances that act on blood vessels are released when tissues are damaged by disease or injury. Histamine is stored by special cells in the skin and elsewhere, and when it is released histamine causes capillary walls to leak fluid, resulting in local tissue-swelling. Prostaglandins and bradykinin have similar functions. All of these substances apparently act locally in the process of inflammation rather than systemically in overall cardiovascular regulation.

Cardiovascular disease. In cardiac failure, cardiac glycosides are used for their inotropic effect on the heart, but vasodilators and drugs that increase urine flow (diuretics; see below *Kidney pharmacology*) are also helpful. The reduced cardiac output resulting from heart failure leads to an increase in pressure in the veins and also to accumulation of tissue fluid (edema). Vasodilators, such as the calcium antagonist nifedipine, dilate the veins and thus lower the venous pressure, and they also increase the cardiac output by reducing the resistance of the arterial system. Diuretic drugs are used to reduce the amount of tissue fluid, but they can also have a beneficial vasodilator effect. These drugs are used in treating high blood pressure as well.

High arterial blood pressure, which is produced by excessive constriction of small arteries, is often treated with drugs. Some drugs that lower blood pressure (hypotensives) inhibit the function of the sympathetic nervous system in various ways (see above *Autonomic nervous system pharmacology*). Hypotensive drugs include methyldopa and clonidine, which probably work at the level of the central nervous system; reserpine and guanethidine, which prevent the release of norepinephrine by sympathetic nerves; adrenoceptor-blocking drugs (*e.g.*, propranolol, which lowers blood pressure by reducing the cardiac output; and prazosin, which blocks the vasoconstrictor action of norepinephrine). Calcium antagonists also have a use in treating hypertension, as do other vasodilators such as hydralazine. A different approach, developed in the late 1970s, consists of using an inhibitor of adrenocortical extract (captopril), thus blocking the formation of angiotensin II. This is very effective in certain types of hypertension in which renin secretion is increased. Most antihypertensive drugs have a variety of unwanted effects, such as drowsiness, dizziness on standing (due to an excessive postural fall in arterial pressure), impotence, and allergic reactions. Though often fairly minor, side effects are a serious problem because of the long-term nature of antihypertensive therapy, and better drugs are constantly being sought.

Treatment of migraine

Migraine is a common condition associated with severe headaches that are believed to result from excessive dilation of the arteries in the membranous covering (meninges) surrounding the brain. The cause is not known, but it is believed to involve the local release of a substance called serotonin. Ergotamine, which comes from a fungus (ergot) that infests cereal crops and has a powerful vasoconstrictor effect, is widely used to treat migraine. Accidental poisoning with the ergot fungus produces many symptoms associated with excessive vasoconstriction, including brain disturbances and gangrene, but they do not generally occur when it is used therapeutically. Other antimigraine drugs include propranolol and calcium antagonists. Tests have shown them to be effective, but it is not clear how they work.

Partial occlusion of the coronary vessels by fatty deposits (atheroma) or blood clots may result in angina pectoris, a pain that occurs when the blood supply to the heart is inadequate for its needs. Vasodilator drugs, particularly nitroglycerine tablets and calcium antagonists, are often used to relieve this condition. They work in part by dilating arteries and veins, which reduces arterial blood pressure and cardiac output, thereby lowering the work and oxygen consumption of the heart. They also have some effect on the coronary vessels themselves and many direct blood toward the regions in which the flow is impaired. Propranolol is also effective because it reduces the rate and force of the heart, thus lowering its oxygen requirement. (H.P.R.)

DRUGS AFFECTING BLOOD

When a small blood vessel is cut, a repair mechanism (hemostasis) is activated that eventually seals the cut and prevents further blood loss. What is in fact a lifesaving mechanism that protects the wounded body from hemorrhage becomes life threatening when clots (thrombi) form within functional blood vessels (thrombosis). Thrombosis tends to occur in blood vessels damaged by artherosclerosis or in vessels with a sluggish blood flow. In veins, portions of the thrombi (emboli) may break off and pass along the bloodstream to become lodged in the arteries of the heart. The drugs described in this section either inhibit hemostasis or they act to enhance the mechanisms that lyse, or dissolve, thrombi.

The clotting process

The clotting process essentially involves the conversion of a soluble plasma protein, fibrinogen, into strands of the insoluble protein fibrin, which forms a mesh that traps platelets. The trigger for hemostasis is an injury to the endothelium, the cells lining the blood vessels, so that the underlying layer of collagen is exposed. The series of events leading to clot formation in a cut blood vessel are (1) constriction of the blood vessel by serotonin, epinephrine, and the thromboxane A_2, which diminishes blood loss; (2) formation of a plug of platelets (the platelet phase) by ADP and thromboxane A_2, also released by platelets, which act in a positive feedback process that makes more platelets adhere to the collagen and to each other; and (3) the conversion of the plug into a clot of fibrin (the coagulation phase). The formation of fibrin entails the sequential interaction of more than a dozen clotting factors, which are protease enzymes (*i.e.*, they accelerate the breakdown of proteins). Each of these clotting factors activates the next in a coagulation cascade of proteolytic reactions that break down protein molecules. The penultimate reaction is the conversion of the soluble fibrinogen to soluble fibrin under the influence of the enzyme thrombin (factor IIa). Soluble fibrin is converted to insoluble fibrin strands by activated factor XIII (fibrin-stabilizing factor), and covalent cross-linkages form between the fibrin strands to give a strong and rigid network. Several of the clotting factors (II, VII, IX, X) require the presence of vitamin K for their activation. Consequently, inhibition of vitamin K blocks the propagation of coagulation pathways.

Under normal conditions the adhesion of platelets to vessel walls is prevented by the vascular endothelial cells, at least in part by their ability to release prostaglandins called prostacyclin or prostaglandin I_2, which reduce platelet stickiness and cause dilation of the blood vessels.

A fibrinolytic system exists in the body that restricts thrombus propagation beyond the site of injury and is also involved in the lysis of clots as wounds heal. The fibrinolytic system degrades fibrin and fibrinogen to products that act to inhibit the enzyme thrombin. The active enzyme involved in the fibrinolytic process is plasmin, which is formed from its precursor, plasminogen, under the influence of an activating factor released from endothelial cells. If formed in the circulating blood, plasmin is normally inhibited by a circulating plasmin inhibitor.

Anticoagulant drugs. Anticoagulant drugs prevent the formation of thrombi by inhibiting the coagulation phase. They are used to prevent the formation and spread of venous and arterial thrombi; however, they are ineffective against existing thrombi. Anticoagulant therapy is used to treat deep-vein thrombosis and pulmonary embolism arising after immobilization or surgery; systemic or coronary arterial embolism caused by heart diseases or replacement of the prosthetic valve; and disseminated intravascular coagulation, which is a systemic activation of the coagulation

system that leads to consumption of coagulation factors and hemorrhage.

Heparin

Heparin, used primarily in hospitalized patients, is a mixture of negatively charged mucopolysaccharides. An endogenous substance whose physiological role is not understood, heparin blocks the coagulation cascade by promoting the interaction of a circulating inhibitor of thrombin (antithrombin III) with activated clotting factors. Because it is not well absorbed when taken orally, heparin is given intravenously to inhibit coagulation immediately; the onset of the drug's effect is delayed after subcutaneous administration. Heparin is not bound to plasma proteins, it is not secreted into breast milk, and it does not cross the placenta. The drug's action is terminated by metabolism in the liver and excretion by the kidney. The major side effect associated with heparin is hemorrhage; thrombocytopenia (reduced number of circulating platelets) and hypersensitivity reactions also occur. Oral anticoagulants and heparin have additional anticoagulant effects. Heparin-induced hemorrhage may be reversed with the antagonist protamine, a positively charged protein that has a high affinity for heparin's negatively charged molecules, thus neutralizing the drug's anticoagulant effect. When given in combination with heparin, dihydroergotamine, which constricts veins and increases blood flow, increases heparin's antithrombotic effect.

Oral anticoagulants

Oral anticoagulants are derivatives of 4-hydroxycoumarin (coumarin) or indan-1,3-dione (indandione). Structurally the coumarin derivatives resemble vitamin K, an important element in the synthesis of a number of clotting factors. Interference in the metabolism of vitamin K in the liver by coumarin derivatives gives rise to clotting factors that are defective and incapable of binding calcium ions (another important element in the activation of coagulation factors at several steps in the coagulation cascade). When anticoagulants are taken orally, several hours are required for the onset of the anticoagulant effect because time is required both for their absorption from the gastrointestinal tract and for the clearance of biologically active clotting factors from the blood. Warfarin, the most commonly used oral anticoagulant, is rapidly and almost completely absorbed; the absorption of dicumarol and other anticoagulant agents, however, is slower and less consistent.

Oral anticoagulants bind extensively to plasma proteins, have relatively long plasma half-lives, and are metabolized by the liver and excreted in the urine and feces. They may cross the placenta to cause fetal abnormalities or hemorrhages in neonates; their appearance in breast milk apparently has no adverse effect on nursing infants. Hemorrhage is the principal toxic effect during oral anticoagulant therapy, but each of the coumarin derivatives causes its own idiosyncratic side effects. Vitamin K, when given intravenously to promote the synthesis of functional clotting factors, stops bleeding after several hours. Plasma that contains normal clotting factors is given to control serious bleeding. Oral anticoagulants may interact adversely with other drugs that bind to plasma proteins or are metabolized by the liver.

Drugs affecting platelets. Platelet aggregates and thrombi formed in coronary arteries may cut off the blood supply to a region of the heart and precipitate a myocardial infarction (heart attack). When administered shortly after a heart attack, drugs affecting platelets can reduce the extent of damage to the heart muscle and the incidence of immediate reinfarction and death. They act in different ways, and the long-term (five-year) benefit of immediate postinfarction therapy is uncertain.

Cyclooxygenase

Aspirin, a non-narcotic analgesic, antipyretic, and anti-inflammatory agent, inhibits an enzyme (cyclooxygenase) involved in the production of thromboxane A_2 in platelets and of prostacyclin in the endothelial cells that line the heart cavities and walls of the blood vessels. Cyclooxygenase is synthesized by endothelial cells but not by platelets. The goal of aspirin therapy is to neutralize cyclooxygenase only in platelets, which inhibits thromboxane A_2 synthesis and therefore platelet aggregation, but to continue the production of cyclooxygenase and prostacyclin in endothelial cells. This is accomplished with a low dose. The occurrence of coronary embolization and the incidence of acute myocardial infarction and death are reduced with the administration of low-dose aspirin therapy. In large aspirin doses, cyclooxygenase synthesis is inhibited in endothelial cells as well as in platelets.

Dipyridamole, a coronary artery vasodilator, decreases platelet adhesiveness to damaged endothelium. The drug prevents platelet aggregation and release by increasing the concentration of platelet cyclic adenosine monophosphate (cAMP) in two ways: by inhibiting an enzyme (phosphodiesterase) that degrades cAMP and by increasing the stimulating effect of prostacyclin on an enzyme (adenylate cyclase) that synthesizes cAMP. Dipyridamole alone does not reduce the incidence of death following myocardial infarction, but it works effectively in combination with other inhibitors of platelet function or with anticoagulants.

Dextran is a plasma volume expander that coats platelets and the blood vessel endothelium, reducing platelet adhesiveness. Dextran reduces the formation of thrombi by increasing their susceptibility to fibrinolysis, and it reduces blood viscosity and dilutes coagulation factors through an osmotic effect. Dextran is similar in effectiveness to heparin and warfarin for preventing venous and pulmonary thromboembolism.

Sulfinpyrazone, a nonsteroidal anti-inflammatory agent, inhibits platelet cyclooxygenase. It has no effect on platelet aggregation but it does prolong platelet revival time. The drug decreases the incidence of postinfarction sudden death, an effect that may be due to an undiscovered aspect of the drug's action unrelated to its effect on platelets.

Drug combinations

Certain drug combinations can be useful for preventing thromboembolism even when the individual agents are without real effect. These combinations include dipyridamole with aspirin or oral anticoagulants after replacement of a prosthetic heart valve; aspirin with dipyridamole after myocardial infarction; and aspirin with sulfinpyrazone after hip surgery.

Fibrinolytic drugs. Fibrinolytic drugs activate the fibrinolytic pathway and lyse clots. The fibrinolytic drugs are distinct from the coumarin derivatives and heparin, which inhibit the formation of clots.

Streptokinase is produced from streptococcal bacteria. When administered systemically, streptokinase lyses acute deep-vein, pulmonary, and arterial thrombi; however, the drug is less effective in treating chronic occlusions. Streptokinase administered by intracoronary artery infusion soon after a coronary occlusion has formed is effective in reestablishing the flow of blood through the heart and vessels after a myocardial infarction and in limiting the size of the area of infarct (or tissue death). Intracoronary infusion permits the delivery of a high concentration of the drug to a localized area and speeds the activation of the fibrinolytic pathway. Intracoronary infusion minimizes the amount of streptokinase inactivated by antibodies that are normally present in blood. Heparin, aspirin, and/or dipyridamole can be added to therapy to help prevent the recurrence of occlusive thrombi. An overdose of streptokinase may lead to bleeding from systemic fibrinogenolysis, which is the breakdown of the coagulation factors by plasmin.

Urokinase is a protease enzyme that activates plasminogen directly. Because it is obtained from tissue culture of human kidney cells, it is not antigenic. Urokinase lyses recently formed pulmonary emboli, and, compared to streptokinase, it produces fibrinolysis without extensive breakdown of the coagulation factors. The usefulness of intravenous or intracoronary urokinase after myocardial infarction is not known.

Tissue plasminogen activator (t-PA) stimulates fibrinolysis, and it has several important advantages over streptokinase and urokinase in treating coronary thrombosis. It binds readily to fibrin and after intravenous administration activates only the plasminogen that is bound to the thrombus; thus, fibrinolysis occurs in the absence of an extensive breakdown of the coagulation factors. It may be used to initiate treatment of heart attack victims en route to the hospital, eliminating the time spent in the hospital preparing the patient for intracoronary injections of streptokinase. This is extremely useful because the rapid reestablishment of coronary blood flow is critically impor-

tant to minimize the amount of damage to myocardial cells after an infarction.

An elevation in the level of circulating plasmin because of excessive activation of the fibrinolytic system may result in fibrinogenolysis and hemorrhage. The antifibrinolytic drug aminocaproic acid is a specific antagonist of plasmin and inhibits the effects of fibrinolytic drugs. (J.S.F.)

DRUGS AFFECTING MUSCLE

Functions of smooth muscle

Smooth muscle. Smooth muscle is found primarily in the internal body organs and performs many functions, including control of the diameter, or calibre, of blood vessels, control of the propulsive activity of the gastrointestinal tract, contraction of the urinary bladder, contraction of the uterus, control of ocular focusing and pupil diameter, and control of the calibre of the respiratory airways. Whereas striated, or skeletal, muscle is controlled from the central nervous system by way of somatic motor nerves, smooth muscle is controlled by the autonomic nervous system and by hormones. In many situations, smooth muscle undergoes spontaneous, often rhythmic, contractions that are not dependent on outside nerve impulses. Smooth muscle contracts much more slowly than striated muscle and in general shows a much broader sensitivity to drugs.

Smooth muscle contraction is initiated by depolarization (the sharp influx of positively charged ions) of the cell membrane. This causes calcium-selective ion channels in the membrane to open, allowing calcium to flow into the cell. The contractile mechanism of smooth muscle cells, like that of striated muscle, involves the sliding action of overlapping protein filaments composed of actin and myosin molecules. The free calcium ions diffuse to the myosin and activate its enzymatic activity, which begins the process of contraction. Most of the drugs that stimulate or inhibit smooth muscle contraction do so by regulating the concentration of intracellular calcium, but other intracellular messengers such as cyclic adenosine monophosphate (cAMP) and cyclic guanosine monophosphate (cGMP) are also involved (see above *General principles*).

The main classes of drugs with important effects on smooth muscle are shown in Table 3. Adrenoceptor agonists, muscarinic agonists, nitrates, and calcium antagonists are considered in other sections and are not discussed here.

Table 3: Drugs Affecting Smooth Muscle

agent	contraction	relaxation
autonomic drugs:		
β-adrenoceptor agonists		GI tract bronchi blood vessels uterus
α-adrenoceptor agonists	blood vessels pupil dilator muscle	
muscarinic agonists	GI tract bladder bronchi pupil constrictor muscle ocular focusing muscle	
local hormones:		
histamine	bronchi GI tract	small blood vessels
bradykinin	bronchi GI tract large blood vessels	small blood vessels
prostaglandins		
PGE series	GI tract uterus	bronchi small blood vessels
PGF series	bronchi uterus GI tract	
oxytocin	uterus milk ducts	blood vessels
drugs:		
nitrates		all smooth muscle, especially large blood vessels
Ca antagonists		blood vessels
ergotamine	uterus blood vessels	
ergometrine	uterus	
papaverine		all smooth muscle
morphine	GI tract (spasm) bronchi	
theophylline		bronchi blood vessels

Several local hormones that are released from cells or formed in tissue act on target cells in close proximity to each other. Because they are destroyed rapidly in the bloodstream they do not function as true blood-borne hormones. Local hormones are usually formed in response to tissue injury, and they are partly responsible for inflammatory and allergic reactions. Smooth muscle responses (*e.g.*, constriction, vasodilation, and edema), particularly in blood vessels and bronchi, are an important component of such reactions. Apart from histamine (see above *Drugs affecting blood*) the main agents known to function as local hormones are kinins and prostanoids.

Kinins

The kinins are peptides that are formed by the enzymatic cleavage of a plasma protein. This cleavage occurs when an enzyme, also present in plasma, is activated in the presence of damaged tissue. Bradykinin, a peptide consisting of a chain of nine animo acids, is an extremely potent vasodilator. Elsewhere in the body, however, bradykinin contracts smooth muscle, particularly in the bronchi and gastrointestinal tract. It also causes the secretion of fluid from the walls of these structures. Constriction of smooth muscle in the bronchi and increased fluid secretion contribute to the airway obstruction that occurs in an asthmatic attack. Bradykinin is probably the causative agent of asthma as well as diarrhea, since similar mechanisms of action occur in the intestine. Bradykinin has no therapeutic uses, but if developed, a selective antagonist of bradykinin might be a useful drug to block its inflammatory and allergic reactions.

Prostanoids (prostaglandins) and leukotrienes (a related group of lipids) are derived by enzymatic synthesis from one of three 20-carbon fatty acids, with the most important in humans being arachidonic acid, a constituent of cell membranes. When a membrane enzyme, phospholipase C_1 is activated, arachidonic acid is released and converted by intracellular enzymes to unstable intermediates, which are further metabolized, depending on the group of enzymes involved, to prostanoids or leukotrienes. The synthesis and release of prostanoids and leukotrienes occurs when cells are damaged, even mildly. They are important in producing tissue responses to injury as well as in other physiological reactions. Derivatives of prostanoids have as their basic structure a five-carbon ring with two side chains, and they differ from each other in the substitutions on the ring structure. The derivatives are distinguished by the letters A through I. In relation to smooth muscle, the most important prostanoids are prostaglandins E_1, E_2, and F_2 (the subscript numbers denoting the 20-carbon precursor and the number of double bonds in the molecule) and leukotrienes C_4 and D_4; the most important sites of action are bronchial and uterine smooth muscle (see Table 3). Leukotrienes are powerful bronchoconstrictors, and they are believed to be synthesized and released during asthmatic attacks. Prostaglandins in minute amounts produce a broad range of physiological effects in almost every system of the human body. Prostaglandins E_1 and E_2 are dilators, and prostaglandins of the F series are bronchoconstrictors. Prostaglandin E_1 also dilates blood vessels, and it is sometimes administered by intravenous infusion to treat peripheral vascular disease. Most prostaglandins cause uterine contraction, and they are sometimes administered to initiate labour (see below *Reproductive system pharmacology*).

Derivatives of ergot

Ergot alkaloids (see above *Cardiovascular system pharmacology*) are produced by a parasitic fungus that grows on cereal crops. Among the many biologically active constituents of ergot, ergotamine and ergometrine are the most important. The main effect of ergotamine is to constrict blood vessels, which can be so intense as to cause gangrene of fingers and toes, giving rise to the name St. Anthony's Fire for the syndrome produced by ergot poisoning. Dihydroergotamine, a derivative, is used in treating migraine (see above *Cardiovascular pharmacology*). Ergometrine has much less effect on blood vessels but a stronger effect on the uterus. It can induce abortion,

though not reliably. Its main use is to promote a strong uterine contraction immediately after parturition, thus reducing the likelihood of bleeding. Both ergotamine and ergometrine cause smooth muscle to contract.

Morphine, an opioid widely used for its painkilling properties, causes smooth muscle contraction in certain situations, which gives rise to some of its side effects. It contracts bronchial smooth muscle (probably by releasing histamine) and may precipitate an attack of asthma. It also causes spasm of the sphincters of the gastrointestinal tract, giving rise to constipation, and spasm of the biliary and urinary tracts. It is therefore not generally suitable for treating pain associated with renal or biliary stones.

Skeletal muscle. Skeletal muscle contracts in response to electrical impulses that are conducted along motor nerve fibres originating in the brain or spinal cord. The motor nerve fibres reach the muscle fibres at sites called motor end plates, located roughly in the middle of each muscle fibre. The motor end plate stores vesicles of the neurotransmitter acetylcholine. An impulse arriving at the motor end plate causes many acetylcholine-containing vesicles to be discharged into the narrow synaptic cleft between the end plate and the membranes of the muscle fibre. Acetylcholine binds to nicotinic receptors on the muscle fibre membrane, causing ion channels to open and allowing a local influx of positively charged ions into the muscle fibre. The muscle fibre is thus depolarized (*i.e.*, its internal potential becomes less negative), and if this local depolarization is large enough a propagated electrical impulse is set up that activates the contractile machinery along the whole length of the fibre. The process occurs within one to two milliseconds (msecs). The released acetylcholine is inactivated within one msec by the action of the enzyme acetylcholinesterase, which is located in the synaptic cleft. The process normally has a large margin of safety because the amount of acetylcholine released is more than enough to activate the muscle fibre.

Drug action on neuromuscular junction

Because the contractile mechanism of skeletal muscles is relatively insensitive to drug action, the most important group of drugs that affect the neuromuscular junction act on (1) acetylcholine synthesis, (2) acetylcholine release, (3) acetylcholine receptors, or (4) acetylcholinesterase.

Hemicholinium and botulinum toxin each cause neuromuscular paralysis by blocking acetylcholine synthesis and acetylcholine release, respectively (see above *Autonomic nervous system pharmacology*). There are a few drugs that facilitate acetylcholine release, including tetraethylammonium and 4-aminopyridine. They work by blocking potassium-selective channels in the nerve membrane, thereby prolonging the electrical impulse in the nerve terminal and increasing the amount of acetylcholine released. This can effectively restore transmission under certain conditions, but these drugs are not selective enough for their actions to be of much use therapeutically.

Neuromuscular blocking drugs act on acetylcholine receptors and fall into two distinct groups: nondepolarizing (competitive) and depolarizing blocking agents.

Competitive neuromuscular blocking drugs act as antagonists at acetylcholine receptors, reducing the effectiveness of acetylcholine in generating an end-plate potential. When the amplitude of the end-plate potential falls below a critical level, it fails to initiate an impulse in the muscle fibre, and transmission is blocked. The most important competitive blocking drug is tubocurarine, which is the active constituent of curare, a drug with a long and romantic history and one of the first drugs whose action was analyzed in physiological terms. Claude Bernard, a 19th-century French physiologist, showed by experiment that curare causes paralysis by blocking transmission between nerve and muscle, without affecting nerve conduction or muscle contraction directly. Curare is a product of plants (mainly *Chondodendron* species) that grow primarily in South America and has been used there for centuries as an arrow poison. Tubocurarine is a complex molecule containing two basic groups that are thought to bind to the receptor in the same way that acetylcholine does.

Anesthetic actions of tubocurarine

Tubocurarine is used in anesthesia to produce the necessary level of muscle relaxation. It is given intravenously, and the paralysis lasts for about 20 minutes, although some muscle weakness remains for a few hours. After it has been given, artificial ventilation is necessary because breathing is paralyzed. Tubocurarine tends to lower blood pressure by blocking transmission at sympathetic ganglia, and, because it can release histamine in tissues, it also may cause constriction of the bronchi. Synthetic drugs are available that have fewer unwanted effects—for example, gallamine and pancuronium. The action of competitive neuromuscular blocking drugs can be reversed by anticholinesterases, which protect acetylcholine against rapid hydrolysis and can increase the amplitude of the end-plate potential enough to restore effective transmission. This is a useful way to restore muscle function at the end of a surgical operation.

Anticholinesterase drugs (see above *Autonomic nervous system pharmacology*) inhibit the rapid destruction of acetylcholine at the neuromuscular junction and thus enhance its action on the muscle fibre. Normally this has little effect, but in the presence of a competitive neuromuscular blocking agent, transmission can be restored. This provides a useful way to terminate paralysis produced by tubocurarine or similar drugs at the end of surgical operations. Neostigmine often is used for this purpose, and atropine is given simultaneously to prevent the parasympathetic effects that are enhanced when acetylcholine acts on muscarinic receptors.

Anticholinesterase drugs also are useful in treating myasthenia gravis, in which progressive neuromuscular paralysis occurs as a result of the formation of antibodies against the acetylcholine receptor protein. The number of functional receptors at the neuromuscular junction becomes reduced to the point where transmission fails. Anticholinesterase drugs are effective in this condition because they enhance the action of acetylcholine and enable transmission to occur in spite of the loss of receptors; they do not affect the underlying disease process. Neostigmine and pyridostigmine are the drugs most often used because they appear to have a greater effect on neuromuscular transmission than on other cholinergic synapses, and this produces fewer unwanted side effects. The immune mechanism responsible for the inappropriate production of antibodies against the acetylcholine receptor is not well understood, but the process can be partly controlled by treatment with steroids or immunosuppressant drugs such as azathioprine.

Depolarizing neuromuscular blocking drugs, of which succinylcholine is the only important example, act in a more complicated way than nondepolarizing, or competitive, agents. Succinylcholine has an action on the end plate similar to that of acetylcholine. When given systemically, it causes a sustained end-plate depolarization, which first stimulates muscle fibres throughout the body, causing generalized muscle twitching. Within a few seconds, however, the maintained depolarization causes the muscle fibres to become inexcitable, so that they fail to respond to nerve stimulation. The paralysis lasts for only a few minutes because the drug is quickly inactivated by cholinesterase in the plasma. Succinylcholine often is used to produce paralysis quickly at the start of a surgical operation (and then is supplemented later with a competitive blocking agent) or for brief surgical procedures. It is used widely, despite a number of disadvantages. Generalized muscle aches are commonly experienced for a day or two after recovery. More seriously, a small proportion of people (about one in 3,000) have abnormal plasma cholinesterase and may remain paralyzed for a long time. Succinylcholine also causes the release of potassium ions from muscles and an increase in the concentration of potassium in the plasma. This happens particularly in patients with severe burns or trauma in whom it can cause potentially dangerous cardiac disturbances. Another hazard is the development of malignant hyperthermia, a sudden rise in body temperature caused by increased tissue metabolism. This condition is very rare, but it is often fatal if not treated rapidly enough.

Paralytic effect of succinylcholine

(H.P.R.)

DIGESTIVE SYSTEM PHARMACOLOGY

Drugs may act on the gastrointestinal tract either by affecting the actions of the involuntary muscle (motility) and thus altering movement or by altering the secretion of

digestive juices. The former drugs may be used to combat diarrhea, constipation, or vomiting.

Drugs affecting gastrointestinal motility. *Diarrhea.* Diarrhea is the frequent passage of a watery, unformed stool. Its causes range from serious organic disease to mental stress. In the treatment of diarrhea, kaolin powder is the most widely used adsorbent powder. Kaolin is a naturally occurring hydrated aluminum silicate, which is prepared for medicinal use as a fine powder. It is suggested that kaolin binds and thus inactivates poisons that may be in the gastrointestinal tract. It is not harmful, and it is effective in many cases if taken in large enough doses (*e.g.*, an initial dose of two to 10 grams followed by the same amount after every bowel movement).

Treatment

Morphine, codeine, and the synthetic opiates and their analogues have a constipating action (morphine was used for this effect long before it was used as a painkiller). The dangers of dependency and addiction clearly prevent the use of certain opiates (*e.g.*, morphine, meperidine, and heroin) as a treatment for diarrhea. Other opiates (codeine and the synthetic analogues diphenoxylate and loperamide) produce little dependence, however, and they have been used successfully in the treatment of diarrhea. Diphenoxylate in combination with atropine has been used by astronauts to reduce bowel movements.

Constipation. There are four kinds of medication available for relief of constipation: saline purgatives, fecal softeners, contact purgatives, and bulk laxatives.

Saline purgatives are salts containing multivalent ions. These highly charged ions do not readily cross cell membranes, and hence they remain inside the lumen, or passageway, of the bowel. They retain water through osmotic forces. The volume of the contents of the bowel is thus increased, stretching the colon and producing a normal stimulus to contraction of the muscle, which leads to defecation. Some commonly used salts are magnesium sulfate (Epsom salts), magnesium hydroxide, sodium sulfate (Glauber's salt), and potassium sodium tartrate (Rochelle salt or Seidlitz powder).

Fecal softeners are not absorbed from the alimentary canal. They act to increase the bulk of the feces. Two are used. The first is liquid paraffin (mineral oil), used either as the oil itself or as a white emulsion. The second group contains drugs with a detergent action. These drugs produce fecal softening in one or two days by increasing the penetration of the stool by water.

Purgatives

Purgatives are drugs whose exact mechanism of action is unknown. They include the anthraquinone derivatives (cascara, aloe, senna, and rhubarb), phenolphthalein, and ricinoleic acid (castor oil). These drugs, frequently used by older people, irritate the lining of the bowel, which may account for their effect. After regular use, their effect tends to lessen, so that larger and more frequent doses are necessary until finally they cease to be effective. They are useful, however, when short-term purging is required (*e.g.*, before surgery or after an illness).

Bulk laxatives act by increasing the size of the feces, in part because of their capacity to attract water. This group includes methylcellulose and carboxymethylcellulose, the gums agar and tragacanth, psyllium (plantago) seed, and dietary fibre.

Vomiting. Emetics produce nausea and vomiting, and their use is limited to the treatment of poisoning with certain toxins that have been swallowed. The most commonly used drugs for this purpose are ipecac syrup, prepared from the dried roots of *Cephaelis ipecacuanha*, a plant indigenous to Brazil and Central America, and apomorphine, a derivative of morphine.

Antiemetics are drugs that prevent vomiting. Broadly, they may be divided into two groups: drugs that are effective in combating motion sickness and drugs that are effective against nausea and vomiting due to other causes. The exact way in which these drugs work is not known. They may act by depressing the chemoreceptor trigger zone in the hypothalamus that controls vomiting.

Drugs that are effective against motion sickness are the anticholinergic drugs and the antihistamines. Although many are available for use, none is entirely free from side effects (*e.g.*, dry mouth and blurred vision with the anticholinergics; drowsiness with the antihistamines). The most effective drugs in this group are the anticholinergic drug scopolamine and the antihistaminic drug promethazine, although many other analogues are used widely.

Nausea and vomiting other than that associated with motion sickness are present in many diseases; *e.g.*, radiation sickness, postoperative vomiting, and liver disease. In these cases, the most effective antiemetics are the phenothiazines (also used in psychiatric medicine) and metoclopramide.

Drugs affecting digestive juices. It has long been thought that gastric and duodenal ulcers are related to the production of acid. Hydrochloric acid is a natural element in the stomach, and excess levels of it, together with the enzyme pepsin, quickly produce ulceration unless the stomach wall is protected. Although investigations have demonstrated that many other factors may play a role in the development of peptic ulcer (*e.g.*, immunological factors and personality type), measures designed to prevent the action of hydrochloric acid on the mucosa is still the cornerstone of therapy. These measures include preventing secretion of acid and neutralization of acid in the stomach by antacids.

Acid secretion. Acid secretion can be prevented by blocking acetylcholine, the neurotransmitter that stimulates acid secretion. The drugs that block the actions of acetylcholine have many side effects (dry mouth, blurred vision, difficulty in passing urine), and although several analogues have been produced only a few have been adequately demonstrated to block gastric acid secretion; they include atropine sulfate and hyoscine. In 1972 a new group of drugs that block gastric acid secretion was discovered; these were called H_2 blockers and include cimetidine and ranitidine (see below *Histamine and antihistamines*). In patients with duodenal ulcers, H_2 blockers administered for four to six weeks dramatically increase the healing of the ulcer and also reduce the pain. Low maintenance doses lower the recurrence rate of duodenal ulcer.

H_2 blockers

Neutralization. Excess acid may be neutralized in the stomach with antacid tablets and mixtures. Antacids do relieve pain, and they have been used since the time of Pliny for this purpose. It has not been proved that they have any effect on the healing of the ulcer. The main constituents of antacids are aluminum and magnesium hydroxides. There are three side effects of antacid therapy, which depend on the compound used. First, many have an action on the bowel; some have a mild laxative effect and some are constipating. Second, if the positively charged compounds are absorbed, the blood may be made alkaline. Third, antacid therapy may affect the absorption of other drugs by binding with them in the gastrointestinal tract.

Mucosal barrier. The mucosal barrier is the name given to the barrier in the stomach that resists the back-diffusion of hydrogen ions. The barrier is a layer of thick mucus secreted together with an alkaline fluid. Since the mucus is a gel, it entraps the alkaline fluid so that the stomach is coated with a tenacious, slimy, alkaline coat. It is probably the integrity of the barrier that is the most important protection the stomach has against attack by acid and pepsin. Many substances attack the mucosal barrier; these include bile, some drugs including salicylates, and substances in the diet. Alcohol has a particularly devastating effect on the barrier, which explains the gastritis (inflammation of the stomach lining) that often follows excess drinking.

There is a group of drugs that have the effect of strengthening the mucosal barrier. These are mainly based on licorice, which has been used in Europe for many years for the treatment of indigestion. Carbenoxolone, a synthetic compound derived from naturally occurring substances in licorice root, significantly increases the rate of healing of peptic ulcers, especially gastric ulcers. The disadvantage of the drug is its side effects; in about one-third of patients it raises the blood pressure and produces fluid retention and potassium loss. These rather serious side effects have led to the development of similar drugs to promote ulcer healing with reduced side effects. (M.A.Su.)

REPRODUCTIVE SYSTEM PHARMACOLOGY

Several sites in the human reproductive system are either vulnerable to chemicals or can be manipulated by drugs.

Within the central nervous system, sensitive sites include the hypothalamus (and adjacent areas of the brain) and the anterior lobe of the pituitary gland. Regions outside the brain that are vulnerable include the gonads (*i.e.,* ovary or testis), the uterus in the female, and the prostate in the male.

Barriers against drugs

The body has anatomical or physiological barriers that tend to protect the reproductive system. The so-called placental barrier and the blood–testis barrier impede certain chemicals, although both allow most fat-soluble chemicals to cross. Drugs that are more water soluble and that possess higher molecular weights tend not to cross either the placental or the blood–testis barrier. In addition, if a drug or chemical binds to a large molecule such as a blood-borne protein, it is less likely to be transported into the testes or less likely to come in contact with the fetus. There appears to be little, if any, barrier to chemicals or drugs gaining entry to breast milk or semen.

Female reproductive system. *Oral contraceptives.* Oral contraceptives, universally known as "the Pill," constitute a class of synthetic steroid hormones. They are capable of suppressing the release of follicle-stimulating hormone (FSH) and luteinizing hormone (LH) from the anterior lobe of the pituitary gland. Known collectively as gonadotrophic hormones, FSH and LH are capable of stimulating the release of progesterone and estrogen from the gonads, or ovaries; all of these hormones are responsible for modulating the female menstrual cycle. Ovulation is believed to be related to a midcycle release of LH, which can be effectively suppressed or blocked by the systematic administration of synthetic hormones. There are many commercial preparations of oral contraceptives, but most of them contain a combination of an estrogen (usually ethinyl estradiol) and a progestin (commonly norethindrone). In general, oral contraceptives are taken in a monthly regimen that parallels the menstrual cycle. Protection from pregnancy is often unreliable until the second or third drug cycle, and during this time certain side effects such as nausea, breast tenderness, or breakthrough bleeding may be evident. More serious side effects, including venous and arterial thromboembolism and a rise in blood pressure, are possible, especially in women over 34 years of age. Normal ovulation usually commences two to three months after stopping the Pill.

Progestin-only preparations (the so-called Minipill) thicken the mucus lining of the cervix and make it more acidic, thereby rendering it hostile to the male spermatozoa. Progestin-only preparations are somewhat less reliable than the combination preparations but produce fewer side effects. Under certain circumstances, the progestin may be administered as an intramuscular deposit that gradually releases the hormone over the course of one to three months.

The Morning-After Pill

Short courses of a high-dose estrogen (the so-called Morning-After Pill) may be taken after coitus. It increases the activity of the fallopian tubes so that the fertilized egg is expelled into the uterus before the uterus has undergone the modification necessary to receive it. This type of contraceptive produces a great deal of nausea and vomiting.

Apart from oral contraceptives, no other drugs are used therapeutically to affect ovarian function. Some drugs may have undesirable side effects on the ovary, which often culminate in menstrual irregularities. Certain tranquilizers (*e.g.,* reserpine and chlorpromazine), narcotics, and anticancer drugs can adversely affect the hormonal secretions of the ovary.

Oxytocin drugs. Oxytocin occurs naturally as a hormone secreted by the posterior lobe of the pituitary gland or it can be made synthetically. Physiologically, it promotes the secretion of breast milk and stimulates the contraction of the uterus during labour. Oxytocin can also be given to control bleeding after childbirth, although one of the ergot alkaloids (*e.g.,* methylergonovine) is more commonly used. Oxytocin can be administered intravenously, sublingually, or by nasal spray.

Teratogenicity. If the fetus is exposed in the uterus to certain environmental chemicals, infections, or drugs, it may develop abnormalities. The toxic substance is described as teratogenic (literally, "monster-producing"), and the study of this type of toxicity is called teratology. About 3 percent of developmental abnormalities have been proved to be drug-induced. It is wise to avoid all drugs (including nicotine) during pregnancy, unless the medicine is well tried and essential. Drugs taken by male partners may be teratogenic if they damage the genetic material (chromosomes) of the spermatozoa.

Male reproductive system. The only male reproductive organ that is a target of pharmacologic manipulation is the prostate gland. This accessory sex organ is susceptible to both benign and malignant changes that seem to be linked to age. The prostate gland is affected by the general class of hormones called androgens, which comprise the male sex hormones. A chemical class of drugs known as antiandrogens is used therapeutically to treat selected pathologic changes in the prostate, which usually include an increase in growth. Antiandrogens can be subdivided into those drugs that contain inherent hormonal properties (*e.g.,* estrogens, or synthetic estrogen such as diethylstilbestrol) and those that possess no hormonal properties (*e.g.,* flutamide). Cyproterone acetate, spironolactone, and Leuprolide are still other examples of antiandrogens or antiandrogen-like agents used in the treatment of prostate disorders.

The androgens

In males there is no process comparable to ovulation and no cyclic release of gonadotropins (hormones that stimulate the gonads, or testes). There is, therefore, no need to suppress the release of gonadotropins by anterior pituitary-hypothalamic axis. Alteration of the amount of pituitary gonadotropins released in the male results largely from the undesirable side effects of various antihypertensive medicines, tranquilizers, and morphine or morphine-like substances. Ultimately such alteration may manifest itself as sterility, impotence, or loss of libido.

The male gonad also does not represent a purposeful target for pharmacologic agents. The testes are, however, affected adversely by the side effects of certain drugs (*e.g.,* anticancer agents) or by exposure to certain occupational or environmental hazards. The most vulnerable type of cells in the testes are the germ cells, which are involved in the process of sperm maturation (spermatogenesis). Immature germ cells (spermatogonia) seem to be the most susceptible to damage by chemicals.

Gossypol, an extract obtained from cottonseed oil, is used as a male oral contraceptive in China. While it has had only limited success, gossypol is an example of a gonadotoxin that cannot inhibit mitotic activity. It produces testes with only Sertoli cells, an action similar to that caused by so-called radiomimetic drugs (*e.g.,* busulfan). (J.A.T.)

KIDNEY PHARMACOLOGY

The kidney is primarily concerned with maintaining the volume and composition of body fluids. It nonselectively filters blood, under pressure, in millions of small units called glomeruli. Large molecules (such as proteins) and cells (such as red blood cells) do not normally pass through the filter into the urine. The filter differentiates only by size, so that useful substances (such as glucose) are filtered out as well as waste products (such as urea, the end-product of nitrogen metabolism). The kidney compensates for this by reabsorbing essential substances and water through the walls of fine tubules, or nephrons, which collect together to deliver their contents into the ureter and then to the bladder, from which urine can be voided. In humans, one litre of filtrate is formed in eight minutes, yet 99 percent of this volume is normally reabsorbed, unless there has been excess fluid intake.

All body fluids have approximately the same strength or tonicity, otherwise considerable osmotic pressures would develop between different compartments. In edema there is excess body fluid with dissolved solutes. When these solutes are absorbed through the walls of the nephrons, after filtration by the glomeruli, an obligatory movement of water, driven by osmotic forces, prevents the body from eliminating excess fluid. By preventing reabsorption of solutes across the walls of the nephrons, both excess solutes and water pass into the bladder. The major use of diuretics is to rid the body of edema fluid that builds up in disease states. Different solutes are moved across the

walls of the nephron by different mechanisms. Diuretics increase the production of urine by interfering with the mechanisms of solute transport.

Regions of the nephron

The nephron can be divided into distinct regions in which the absorptive processes are different: the proximal tubule, leading directly from the glomerulus; the loop of Henle; the distal tubule, leading away from the loop; and the collecting duct. The majority of useful solutes and water are reabsorbed in the proximal tubule, while selective absorption, regulation, and fine tuning to maintain the composition of body fluids in the correct ranges take place in the remaining regions.

Proximal tubule. Carbonic anhydrase inhibitors, such as acetazolamide, dichlorphenamide, and methazolamide, depress the reabsorption of sodium bicarbonate in the proximal tubule by inhibiting an enzyme, carbonic anhydrase, which is involved in the reasborption of bicarbonate. Urine formation is increased. The urine, which is rich in sodium bicarbonate and is alkaline, also has an increased concentration of potassium ions, which can lead to a serious loss of potassium from the body.

The loop of Henle. Diuretics that act in the loop of Henle produce a rapid peak in the excretion of urine (diuresis), which then wanes as the drugs are excreted and because of the compensatory factors due to fluid loss. These diuretics clear sodium chloride (salt) from the body and interfere indirectly with the mechanisms by which water is reabsorbed from the collecting duct. Consequently, large volumes of dilute urine containing sodium, potassium, and chloride ions are formed. The loop diuretics are often also called "high-ceiling diuretics" because they can produce an extra level of diuresis over and above the maximum produced by other classes of diuretic drugs. Examples of this class are furosemide, ethacrynic acid, piretanide, and bumetanide.

Distal tubule. The thiazide class of diuretics interferes with salt reabsorption in the first part of the distal tubule. A mild diuresis results in which sodium, potassium, and chloride ions are eliminated in the urine. The thiazides are widely used in the treatment of hypertension. Examples are chlorothiazide, hydrochlorothiazide, and hydroflumethiazide. Chemically, they are related to the carbonic anhydrase inhibitors.

Role of aldosterone

The adrenal gland releases a hormone, aldosterone, which promotes sodium absorption in the latter part of the distal tubule. Its function is to increase sodium retention in sodium-depleted states. Aldosterone levels, however, may be abnormally high in hyperaldosteronism and in hypertension. Drugs such as spironolactone act as antagonists of aldosterone and compete with it for its site of action in the distal tubule. As with most antagonists, spironolactone has no direct action of its own but simply prevents the action of the hormone, thereby correcting the excess sodium reabsorption.

In the latter part of the distal tubule there are mechanisms that exchange one ion for another; for example, sodium is exchanged for potassium and sodium is exchanged for hydrogen. Sodium is absorbed across the tubule wall while potassium and hydrogen are added to the urine. Thus diuretics such as the thiazides, loop diuretics, and carbonic anhydrase inhibitors, which prevent sodium absorption in the early parts of the nephron, cause an unusually large sodium load to be delivered to the distal tubule. There sodium may be exchanged for other ions, especially potassium, and reabsorbed from the urine. The result is that the body loses a large amount of potassium ions, which is serious if the loss exceeds the capacity of the diet to restore it. Potassium depletion leads to failure of neuromuscular function and to abnormalities of the heart, among other serious effects. The potassium-sparing diuretics amiloride and triamterene block the exchange processes in the distal tubule, and thus prevent potassium loss. Sometimes a mixture of diuretics is used in which a thiazide is taken together with a potassium-sparing diuretic to prevent excess potassium loss. In other instances, the potassium loss may be made up by taking oral potassium supplements in the form of potassium chloride.

Osmotic diuretics (*e.g.,* mannitol) are substances that have a low molecular weight and are filtered through the glomerulus. They limit the reabsorption of water in the tubule. Osmotic diuretics cannot be reabsorbed from the urine and so they set up a situation of nonequilibrium across the tubule membrane. In order to maintain normal osmotic pressure, water is moved across the membrane, increasing the volume of urine.

Properties of osmotic diuretics

In some situations it is desirable to change the acidity or alkalinity of the urine, usually to promote the loss of toxic substances from the body. Urine may be made more alkaline by giving sodium bicarbonate or citrate salts. It may be made more acid by giving ammonium chloride.

(A.W.C.)

DERMATOLOGICAL PHARMACOLOGY

The skin consists of layers called the epidermis and the dermis and of certain appendages such as sweat glands, sebaceous glands (which secrete an oily substance), hair, and nails. There also exists a subcutaneous layer beneath the dermis. The outermost layer of the epidermis is termed the stratum corneum. It consists principally of dead epithelial cells that are filled with a protein, keratin, which waterproofs and toughens the skin. Underlying the stratum corneum are layers comprising granular spinous cells, keratinocytes, melanocytes, and Langhan's cells. The dermis, which is below the epidermis, comprises connective tissue and a number of different cell types; it maintains and nourishes the epidermis through its network of capillaries and lymphatic vessels. Sweat glands and hair follicles, which originate in the dermis and penetrate the stratum corneum of the epidermis, represent a potential route of penetration by drugs or chemicals. The subcutaneous layer is the innermost layer and is composed of loose connective tissue and many fat cells. It provides some degree of insulation and is a location for food storage and the site for subcutaneous injection.

Few chemicals or drugs are absorbed rapidly from intact skin. In fact, the skin effectively retards the diffusion and evaporation even of water except through the sweat glands. There are, however, a few notable exceptions (*e.g.,* certain types of nerve gases, as well as insecticides, scopolamine, and nitroglycerin), and instances where a penetration enhancer (*e.g.,* dimethyl sulfoxide) serves as a vehicle for the drug.

Several factors affect the transport of drugs through the skin (percutaneous) once they are applied topically. The absorption of drugs through the skin is enhanced if the drug is highly soluble in the fats (lipids) of the subcutaneous layer. The addition of water (hydration) to the stratum corneum greatly enhances the percutaneous transport of corticosteroids (anti-inflammatory steroids) and certain other topically applied agents. Hydration can be effected by wrapping the appropriate part of the body with plastic film, thereby facilitating dermal absorption. If the epithelial layer has been removed (denuded) by abrasion or burns, or if it has been affected by a disease, penetration of the drug may proceed more rapidly. A drug will be distributed (partitioned) between the solvent and the lipids of the skin according to the solubility of the solvent in water or lipids. Topical absorption of drugs is facilitated when they are dissolved in solvents that are soluble in both water and lipids. Highly water-soluble, polar molecules, which have a lesser tendency to solubility in lipids, essentially cannot be absorbed percutaneously. Thus, a drug penetrates the skin at a rate determined primarily by its tendency to dissolve in water, or lipids, or both.

Topically applied drugs. Topical application of drugs provides a direct, localized effect on a specific area of the skin. When drugs are applied topically to the skin, they may be dissolved in a variety of vehicles or formulations, ranging from simple solutions to greasy ointments. The particular type of dermal formulation used (powder, ointment, etc.) depends, in part, on the type of skin lesion or disease process.

Therapeutic effects of topical application of drugs

Topically applied medications can relieve itching, exert a constricting or astringent action on the pores, or dissolve or remove the epidermal layers. Other pharmacologic effects from topically applied drugs include antibacterial, anti-inflammatory, antifungal, and antiparasitic actions. Analgesic balms (*e.g.,* wintergreen oil or methyl salicy-

late) have been used topically to relieve minor muscle aches and pains.

Steroids. Corticosteroids (anti-inflammatory steroids) have been used for many years to treat dermatologic disorders. The percutaneous absorption of topically applied corticosteroids is facilitated by encasing the afflicted area with a plastic film (*i.e.,* hydration). Triamcinolone, fluocinolone, dexamethasone, methylprednisolone, and hydrocortisone are examples of synthetic corticosteroids that can be applied topically in ointments, creams, or lotions. Prolonged or excessive use of corticosteroids can lead to systemic levels that produce undesirable side effects, such as salt retention or suppression of the function of the pituitary and adrenal glands.

Anticancer agents. Certain anticancer (cytotoxic) agents or immunosuppressant drugs are applied transdermally to skin disorders, including some skin tumours. An example of such an anticancer agent is 5-fluorouracil.

Photosensitizing. The skin can be affected by other means, including sunscreens, photosensitizing drugs, and pigmenting agents (psoralens). Sunscreens, which act as barriers to sunlight by blocking, scattering, or otherwise reflecting the light, include such agents as para-aminobenzoic acid. Other chemicals (*e.g.,* coal tar) act in conjunction with sunlight on the skin to achieve a high sensitivity to sunlight (photosensitization). Drugs capable of causing photosensitization generally exert their effects following the absorption of light energy. For example, the topical or systemic administration of methoxsalen or trioxsalen prior to the exposure to the ultraviolet radiation of the Sun augments the production of melanin pigment in the skin. These and other psoralens have been used in the treatment of the skin disorder vitiligo in an effort to repigment the whitish patches that commonly occur on the hands and face.

Systemic effect of transdermal application

Transdermally applied drugs. The transdermal application of drugs provides an alternate method for applying drugs to achieve a systemic, rather than local, effect. The administration of a drug through the skin not only minimizes the metabolism of the drug before it reaches the rest of the body but also eliminates the high and low blood levels associated with oral administration. A major limitation of transdermal drug administration is that only a small amount of drug can be given through the skin.

Transdermal drug administration makes use of a variety of structures from which the drug is distributed. The rate of drug release is determined by the properties of the synthetic membrane of the vehicle and the difference in drug concentration across the membrane. To ensure that the drug delivery rate remains constant, the skin must be capable of removing the drug at a rate faster than it is released from the device. Because the anatomical site can influence this rate, testing for the most suitable areas is done for each drug. Examples of transdermal drugs are nitroglycerin, in impregnated disks applied to the upper chest or upper arm, and scopolamine (a drug used to treat motion sickness and nausea), applied in a polymer device behind the ear.

Mucous membranes. Drugs may be applied to mucous membranes, including those of the conjunctiva, mouth, nasopharynx, vagina, colon, rectum, urethra, and bladder. They may either exert a local action or be absorbed into the bloodstream to act elsewhere. Examples include nitroglycerin, which is absorbed from under the tongue (sublingually) to act on the heart and relieve anginal pain, and trifluoperazine, which is a tranquilizer sometimes taken in suppositories. Nasal insufflation, or inhalation, involves the local application of a drug to the mucous membranes of the nose to achieve a systemic action. This represents an effective delivery route of antidiuretic hormone and its analogues in the treatment of diabetes insipidus. Relatively unsuccessful efforts have been made to get hormones of larger molecular weight, such as insulin or growth hormone, to penetrate the mucous membranes of the nasal cavity, thereby avoiding the need to inject such hormones. Although certain medications can be applied successfully to mucous membranes, the topical application of drugs to the skin represents a more widespread and important therapeutic method of administration.

Antibiotics. Although certain antibiotics are used topically, their use should be restricted to the most superficial skin infections. Antibiotics are more often administered systemically. The tetracyclines have been used topically for the treatment of acne. Skin disorders caused by fungi can be treated with either antibiotics or antifungal drugs.

ENDOCRINE PHARMACOLOGY

Coordinated control of body functions

Control of most body functions is achieved by the nervous system and the endocrine system, which constitute the two main communication systems of the body. They function in a closely coordinated way, each being dependent on the other for its proper operation; the total behaviour of the organism is integrated by a constant traffic of both neural and hormonal signals, which are received and responded to by appropriate tissues. The activities of the central nervous system and of the endocrine glands are themselves dependent on feedback control through neural and hormonal stimuli. This control is related to the toxicity of hormones when used therapeutically because prolonged use of certain hormones or their analogues in this way may quell, sometimes irreversibly, the appropriate gland's output of endogenous hormone.

The natural hormones belong to only a few chemical classes. Most are polypeptides, some are derivatives of amino acids (epinephrine, norepinephrine, dopamine, or thyroid hormones), and some are steroids (the sex hormones and the hormones of the adrenal cortex). Polypeptide and amino acid hormones bring about their effects by acting on cell membrane receptors that are specifically sensitive to their action. Steroid hormones penetrate the cell membrane and interact with receptors on specific binding proteins, which then act on the cell nucleus to modify protein synthesis. The techniques of recombinant DNA technology have begun to provide improved methods for obtaining large amounts of scarce human hormones in pure form.

The functions of hormones fall into three general categories: (1) morphogenesis, which is a process that uses hormones to regulate the growth, differentiation, and maturation of the organism (*e.g.,* the development of secondary sex characteristics under the influence of ovarian or testicular hormones); (2) homeostasis, or metabolic regulation, in which hormones are used to maintain a dynamic equilibrium of the components of the body, such as fats, carbohydrates, proteins, electrolytes, and water; and (3) functional integration, whereby hormones regulate or reinforce functions of the nervous system and patterns of behaviour (*e.g.,* the influence of sex hormones on sexual activity and maternal behaviour).

The endocrine system comprises the anterior and posterior lobes of the pituitary gland, the adrenal gland, the pancreas, the gonads (ovaries and testes), and the thyroid and the parathyroid glands. Several of the endocrine glands (thyroid, adrenal cortex, ovaries, and testes) are under the control of the hormones of the anterior lobe of the pituitary gland. The hormones that are released to control the actions of other hormones are called trophic hormones. The release of trophic hormones is under the control of neurons of the hypothalamus, and it is mainly at this level that integration of the activities of the nervous and endocrine systems takes place.

Trophic hormones

The therapeutic use of hormones is concerned primarily with replacement therapy in deficiency states (*e.g.,* deficiency of glucocorticoids in Addison's disease). Hormones and their analogues and antagonists, however, can be used for a variety of additional purposes—*e.g.,* topical corticosteroids and oral contraceptives.

Anterior pituitary gland. The pituitary gland (also called the hypophysis) is situated at the base of the brain. The gland itself is composed of three distinct lobes: the anterior lobe (also called the adenohypophysis), the posterior lobe (also called the neurohypophysis), and the intermediate lobe (or pars intermedia). The pituitary gland is connected by a bridge, the pituitary stalk, through which it receives its blood supply and many neurohumoral and hormonal signals from the region of the brain known as the hypothalamus. Neurohumours are chemical signals produced by nerve cells, and hormones are produced by

endocrine glands; both act in different ways to affect the endocrine system. Hormones emanating from the hypothalamus, called hypothalamic-releasing hormones, affect the secretion or release of hormones stored in the anterior lobe of the pituitary. For each anterior pituitary hormone there is an appropriate corresponding hypothalamic-releasing hormone. There are six important trophic hormones of the mammalian anterior pituitary gland: growth hormone (GH, also called somatotrophin); prolactin; thyrotropin (thyroid-stimulating hormones, or TSH); adrenocorticotropin (adrenocorticotrophic hormone, or ACTH); luteinizing hormone (LH); and follicle-stimulating hormone (FSH).

Growth hormone. As its name implies, growth hormone stimulates the growth of cells in the body. It does not act on a specific group of cells or organs but rather on all of the cells of the body to promote their growth and proliferation. Growth hormone is a protein hormone whose molecular structure consists of a 191-amino-acid sequence. It stimulates bone growth; hence, it plays an important role in the growing child. Pituitary tumours can sometimes result in oversecretion of GH, leading to gigantism or acromegaly. In the United Kingdom, drug treatment for pituitary tumours is carried out with bromocriptine, a substance that resembles dopamine in its actions. Bromocriptine decreases the overproduction of growth hormone by the tumour cells. The consequence of growth hormone deficiency in children is dwarfism, and it is treated by replacement therapy with human growth hormone produced by recombinant DNA technology.

Effects of pituitary tumours

Prolactin. Prolactin, along with other hormones (*e.g.*, oxytocin), acts on cells of the breast to stimulate growth and enhance the secretion of milk (lactation). Prolactin mediates these hormonal actions through receptors that are located within the mammary glands. Inappropriate or excessive secretion of prolactin may be suppressed by treatment with bromocriptine.

Thyrotropin. Thyrotropin (thyroid-stimulating hormone, or TSH) is secreted by the anterior lobe of the pituitary gland upon the command of thyrotropin-releasing hormone (TRH). Through receptors located in the thyroid gland, TSH stimulates the biosynthesis of a thyroid hormone, thyroxine, and other iodine-containing precursors. There is feedback among TRH, TSH, and thyroxine: if TSH causes the thyroid gland to manufacture too much thyroxine, then thyroxine can travel to the pituitary gland and act on receptors that slow down the secretion of TSH and hence TRH. This negative feedback by thyroxine contributes to the body's ability to maintain appropriate levels of the hormones.

Adrenocorticotropin. Adrenocorticotropin (ACTH), a peptide chain of 39 amino acids, is the smallest of the hormones of the anterior pituitary. Adrenocorticotropin is under the central control of corticotropin-releasing factor (CRF). Adrenocorticotropin stimulates the cortex of the adrenal gland to produce a variety of corticosteroids that affect electrolyte and water balance (mineralocorticoid) or carbohydrate, fat, and protein metabolism (glucocorticoids). The principal hormonal action of ACTH is to stimulate steroid biosynthesis, especially cortisol. It is used mainly to verify the diagnosis of adrenal insufficiency (diminished glandular output), but occasionally it is used in nonendocrine disorders that show some response to glucocorticoids (*e.g.*, multiple sclerosis). Tetracosactrin is the polypeptide consisting of the first 24 amino acids of ACTH. It is often used in preference to ACTH because it produces fewer side effects.

Luteinizing hormone. Luteinizing hormone (LH) is one of the two anterior pituitary hormones collectively referred to as gonadotropins (hormones that regulate the functions of the gonads). The other gonadotropin is follicle-stimulating hormone. While both hormones have specific physiologic actions, they often act in concert to moderate the reproductive processes. In females, LH induces ovulation and maintains secretory functions within the uterus. In males, LH acts to stimulate the secretion of male sex steroids (androgens) by specialized cells in the testes (interstitial cells, or Leydig cells). In females, the pituitary secretion of LH is cyclic and along with other hormones is responsible for the menstrual cycle. In males, other than some minor fluctuations in hormonal levels due to daily variations, the pituitary secretion of LH is steady.

Follicle-stimulating hormone. Like LH, follicle-stimulating hormone (FSH) is a glycoprotein secreted by the anterior pituitary gland. In females, FSH stimulates the development of ovarian follicles as well as the biosynthesis of female sex steroids (estrogens). The rupture of the follicle, thought to be assisted by LH, gives rise to the release of an ovum, or egg. The process is referred to as ovulation. It is necessary that FSH prime the follicles prior to the actual rupturing process caused by LH. Although LH is known as the ovulatory hormone, it could not accomplish this physiologic function without the prior actions of FSH. In males, FSH stimulates two types of specialized cells also found in the testes: primordial germ cells (gametes) and Sertoli's cells (also called sustentacular cells). The maturation of germ cells (spermatogonia) is stimulated by FSH; the process is termed spermatogenesis. Follicle-stimulating hormone also stimulates the secretion of certain proteins in the testes.

Role of FSH in ovulation

Gonadotropins (LH and FSH), extracted from the urine of postmenopausal women, and chorionic gonadotropins (produced by the placenta), extracted from the urine of pregnant women, are used to promote ovulation in infertile women in whom the pituitary is primarily the cause of infertility. Chorionic gonadotropin is sometimes used in the condition called cryptorchidism, in which the testes failed to descend into the scrotum in childhood.

Posterior pituitary gland. The posterior pituitary gland secretes two hormones, oxytocin and vasopressin. Vasopressin is also called antidiuretic hormone (ADH) since one of its physiologic actions is exerted on the kidney, leading to a reduction in urinary output. Oxytocin and ADH are octapeptides whose secretion is modulated by secretory activities of nerve cells (neurosecretion) located in specialized regions of the hypothalamus.

Vasopressin. A disease known as diabetes insipidus is characterized by the excessive production of urine with a high concentration of water, which is the result of the failure of the kidney tubules to reabsorb the proper amount of water. A common cause of diabetes insipidus is the inadequate production of vasopressin by the pituitary. The condition may be treated by replacement of vasopressin or by the use of its synthetic analogue desmopressin. Vasopressin also plays a role as a neurotransmitter in the brain, and there is some evidence that it is involved in memory storage and in mood.

Oxytocin. Oxytocin exerts its hormonal effects on the uterus and the mammary gland. Oxytocin increases the frequency and strength of the concentrations of the muscles of the uterus, particularly during labour. Oxytocin facilitates the birth process by aiding the expulsion of the fetus and the placenta. This hormone also facilitates lactation by stimulating the contraction of myoepithelial cells surrounding the ductal system of the mammary glands. Oxytocin may be used to induce labour, especially when gestation approaches or exceeds 40 weeks.

Adrenal gland. The adrenal gland is a compound organ situated on top of the kidney. It consists of an outer cortex (adrenal cortex) and an inner medulla (adrenal medulla). The hormones secreted from the cortex are steroids, generally classified as mineralocorticoids and glucocorticoids. Those substances emanating from the medulla are amines such as epinephrine and norepinephrine. Epinephrine and norepinephrine are catecholamines, which are present in several cell types in the body and serve as neurotransmitters both in the brain and in the autonomic nervous system. Thus, the adrenal medulla functions with the sympathetic nervous system.

Steroids and amines

The adrenal cortex secretes glucocorticoids (*e.g.*, cortisol) and mineralocorticoids (*e.g.*, aldosterone, which causes sodium retention and potassium excretion by the kidney). Glucocorticoids together with mineralocorticoids are used in replacement therapy in acute or chronic adrenal insufficiency (Addison's disease).

Glucocorticoids, including a range of synthetic analogues (*e.g.*, prednisolone, triamcinolone, and dexamethasone), are used as anti-inflammatory and immunosuppressant

agents. As anti-inflammatory agents they are used topically and in bronchial asthma. Glucocorticoids stimulate the synthesis of a glycoprotein called macrocortin that inhibits the enzyme phospholipase A_2. Phospholipase A_2 plays an essential role in the synthesis of prostaglandins and leukotrienes; its inhibition by macrocortin underlies the anti-inflammatory effects of glucocorticoids.

Glucocorticoids also have an antivitamin D action and accordingly are used in hypercalcemia (excess concentration of calcium in the blood) associated with sarcoidosis or vitamin D overdosage. They are used in some life-threatening diseases, including certain cancers.

Thyroid and parathyroid glands. Anatomically, there is close proximity between the thyroid and the parathyroid gland, yet they have very different hormonal functions and different mechanisms of stimulation. Secretion of thyroxine and triiodothyronine from the thyroid gland is stimulated by TSH, whereas secretion of parathyroid hormone (PTH) from the parathyroid gland and calcitonin from the thyroid gland are modulated by circulating levels of calcium and phosphate in the blood.

Parathyroid hormone. Parathyroid hormone (PTH) is a single-chain amino acid molecule secreted by the chief cells of the parathyroid gland. The major regulator of PTH is the level of ionized calcium in the blood. Receptors for PTH are located in bones and in the kidney. The primary function of PTH is to prevent hypocalcemia (lowered blood calcium levels) by stimulating the release of calcium from bone and by increasing the reabsorption of calcium in the kidney. Tetany, often characterized by muscle twitching, spasms, and even convulsions, is the result of precariously low levels of blood calcium. Tetany resulting from hypocalcemia can be treated by administering calcium gluconate and vitamin D.

Calcitonin. Calcitonin is synthesized in the thyroid, the parathyroid, and the thymus glands. It is capable of counteracting calcium-releasing effects of PTH because calcitonin reduces blood calcium levels by increasing its deposition in bone and enhancing its excretion in the urine.

Thyroxine and triiodothyronine. Thyroxine (T_4) and triiodothyronine (T_3) are the major iodine-containing hormones synthesized in the thyroid gland, and their synthesis is stimulated by pituitary TSH. Iodine is also required in their manufacture. The biochemical actions of the thyroid hormones are complex, although their most evident effect is stimulation of cellular metabolism. Both T_4 and T_3 increase the basal metabolic rate (measure of the minimum number of calories needed to sustain only the activities essential to life) and are calorigenic (energy-producing). A diet deficient in iodide can lead to the formation of a goitre (an enlargement of the thyroid); iodized table salt can prevent this abnormality. Enlargement of the thyroid gland often is due to the unchecked secretion of TSH when there is too little iodide to provide for the manufacture of T_4 or T_3, which are the hormones that normally reduce the secretion of the pituitary trophic hormone TSH. Excessive thyroid hormone production (hyperthyroidism) in adults is commonly referred to as thyrotoxicosis; a common type in young adults is exophthalmic goitre (Graves' disease). In this disease there is increased activity of thyroid hormones as a consequence of an immunoglobulin called long-acting thyroid stimulator (LATS) that is produced in lymphoid tissue. The circulating level of TSH is low because of negative feedback, but LATS is not susceptible to negative feedback. In older patients, thyroid cancers are more commonly the cause of hyperthyroidism. Treatments include surgery; the use of radioactive iodine, which is taken up by the thyroid and then irradiates it; and antithyroid drugs such as carbimazole.

Importance of iodine

Thyroid deficiency in newborns can lead to cretinism, which is characterized by mental retardation. In adults, hypothyroidism leads to the condition called myxedema. Thyroxine is used to treat hypothyroid conditions in both infants and adults.

Pancreas. The pancreas has both an endocrine (secretion of insulin and glucagon) and a digestive function. A deficiency in the pancreatic secretion of insulin leads to diabetes mellitus ("sugar diabetes"). One of insulin's important physiological actions is to control blood sugar (glucose) levels. This carbohydrate is an important nutrient for cellular metabolism, and the cell must receive neither too little nor too much. Diabetes mellitus is a complex metabolic disease that is caused both by genetic factors (juvenile-type diabetes, also called type I diabetes) and by the aging process (maturity-onset, or type II, diabetes) in the pancreatic cells responsible for secreting the protein hormone insulin. The islets of Langerhans in the pancreas contain a specialized type of cell called the *beta,* or B, cell, which secretes insulin. A lack or absence of B cells, among other disease factors, can lead to diabetes mellitus. Once secreted by the B cell into the bloodstream, insulin can affect a number of important metabolic actions on cells located in the muscles, liver, and other sites. Normally, insulin secretion is increased following the ingestion of carbohydrates; the liver is responsible for eventually curtailing the biologic actions of insulin. Insulin has a number of important metabolic actions upon both fat and protein as well.

Insulin

Because insulin is a polypeptide it cannot be administered orally since proteolytic enzymes present in the stomach and gastrointestinal tract destroy its physiological properties. Insulin must be injected parenterally so that it enters the bloodstream and eventually reaches the body's cells. Insulin can be obtained from the pancreas of domestic animals, and the hormone now can be made by bacteria following recombinant DNA techniques. An overdose of insulin can produce hypoglycemia (low blood sugar), which may lead to convulsions. Several other hormones (*e.g.,* growth hormone and glucocorticoids) can antagonize insulin's actions.

Maturity-onset diabetes often may be treated with oral hypoglycemic drugs instead of with insulin. These drugs are of two types. In the United States, only the sulfonylureas (*e.g.,* tolbutamide) are used, which increase the release of endogenous insulin and increase the number of insulin receptors. In the United Kingdom, the biguanides (*e.g.,* metformin) are used to lower blood sugar by uncertain mechanisms that may include decreased absorption, decreased synthesis from protein, increased glucose utilization, and decreased appetite.

Another set of specialized cells located in the pancreas secrete a protein hormone called glucagon. Glucagon can stimulate the breakdown of liver glycogen, leading to a release of glucose into the bloodstream. Thus, under certain instances, glucagon can counteract the actions of insulin. The physiological or pathological significance of the antagonistic relationship between insulin and glucagon is not fully understood. Unlike insulin, there is no known human disease associated with either increased or decreased physiological levels of glucagon. (J.A.T.)

HISTAMINE AND ANTIHISTAMINES

Histamine is a chemical messenger involved in a number of complex biologic actions. It is widely distributed in the plant and animal kingdoms. In humans it occurs mainly in an inactive bound form in most body tissues. When released, it interacts with specific macromolecules (histamine receptors) on the cell surface or within a target cell to elicit changes in many different bodily functions. Histamine is a small, polar, organic molecule of low molecular weight, which comprises an imidazole ring and an ethylamine side chain; it is water soluble and is a base.

Antihistamines are a group of synthetic drugs that can inhibit various actions of histamine. They have some chemical resemblance to histamine and act as antagonists by competing with histamine for occupation of its receptor sites, thereby preventing histamine from eliciting its usual responses. They are helpful therapeutically in preventing rather than in reversing histamine actions.

Antagonistic properties of antihistamines

The chief sites of histamine in the body are the mast cells of connective tissue and their circulating counterparts in the blood, the basophils. These cells synthesize histamine by the action of an enzyme that removes the carboxyl group from the amino acid L-histidine. Histamine is then stored in many tissues of the body. If the histamine is released, more of it is slowly synthesized at these sites. There are, however, organs in which the histamine-

containing cells have not been identified. Some tissues have a high capacity for synthesizing histamine without storing it. Cells in regenerating or rapidly growing tissues also produce large amounts of histamine, which is continuously released.

Free histamine produces many powerful and varied biologic actions. It appears to act on specific receptors in the membranes of cell surfaces. These receptors have not been isolated or identified, but their presence is inferred by the use of synthetic drugs. Three types of pharmacological histamine receptor have been described, and they are designated as H_1, H_2, and H_3.

Histamine stimulates many smooth muscles to contract, such as those in the gastrointestinal tract, the uterus, and the bronchi. In some smooth muscle, however, it causes relaxation, notably that of fine blood vessels, whose dilation may produce a pronounced fall in blood pressure. Histamine also increases the permeability of the walls of the capillaries so that more of the constituents of the plasma can escape into the tissue spaces, leading to an increase in the flow of lymph and its protein content and to the formation of edema. These effects are manifested in the redness and weal associated with histamine release, as may occur after a scratch from a blunt instrument or a nettle sting. There are striking differences, however, in the response of different animal species to histamine. For example, the rat is relatively resistant whereas the guinea pig and man are very sensitive.

Location of histamine

Histamine appears to have a physiological role in the body's defenses against a hostile environment. Histamine is found in the body's surfaces: in the skin, in the respiratory membrane and adjacent tissue, and in the lining of the alimentary, or digestive, canal. Histamine may be released from tissue mast cells and blood basophils when the body is subjected to mechanical damage, burning, infection, or some drugs. Histamine assists the body in removing the products of cell damage from inflammation. In humans, the most common circumstance in which histamine is liberated is as a result of the antibodies produced by foreign proteins. Under extreme circumstances, the effects of histamine become pathological, leading to exaggerated responses with distressing results, as may occur in some allergic conditions.

Synthetic drugs known as antihistamines (*e.g.*, mepyramine, diphenhydramine, and chlorpheniramine) have been available since 1945, although subsequently they have been designated more precisely as H_1-receptor histamine antagonists or H_1-receptor blockers. The H_1 antihistamines are used to suppress or alleviate the symptoms in various allergic conditions; they do so by competing with the released histamine for occupation of its H_1 receptors. They may be effective in the treatment of seasonal hay fever (seasonal rhinitis and conjuctivitis) to relieve sneezing, rhinorrhea, and itching of eyes, nose, and throat. In general, the H_1 antihistamines tend to be more successful in controlling acute than chronic conditions; thus, they are most useful at the beginning of the hay-fever season when the allergens are present in low concentration, but in perennial vasomotor rhinitis (nonseasonal, nonallergic inflammation of the mucous membranes of the nose brought on by environmental or emotional stimuli) they are only of limited value. They are not usually effective in asthma, indicating that in this condition histamine is not the main agent producing the symptoms. Certain allergic skin reactions respond favourably to H_1 antihistamines, which are particularly effective for treatment of acute urticarial rashes (or weals) of the skin and the itch and swelling of insect bites.

The H_1 antihistamines are relatively free from serious side effects, and the margin between therapeutic and toxic dose is generally large. Less serious side effects are common, the most notable being drowsiness. New H_1 antihistamines, however, are relatively free of this side effect. The action of some H_1 antihistamines on the central nervous system has been put to advantage in the prevention and treatment of motion sickness and nausea and of mild insomnia. The main antihistamines used in the treatment of motion sickness include cinnarizine, cyclizine, dimenhydrinate, mepyramine, and promethazine. Because they also possess sedative action (especially dimenhydrinate and promethazine), they may impair a person's performance while driving and enhance the effects of alcohol and other depressant drugs that act on the central nervous system. Antihistamines bind strongly to H_1 receptors in the brain, but it is not known whether this action is responsible for their beneficial effect in motion sickness. Some also bind to mascarinic receptors in the brain, and this action may contribute to their beneficial effect.

Histamine has a physiological role in regulating the secretion of acid in the stomach, where it stimulates the parietal cells to produce hydrochloric acid. This is probably protective, since the acid controls the local bacterial population. A pathological situation can arise, however, as in the formation of gastric or duodenal ulcers. In the 1970s a new class of synthetic drugs was invented that blocked the action of histamine at its H_2 receptors. These drugs were shown to be extremely effective in antagonizing the action of histamine in stimulating acid secretion and in blocking other stimulants of acid secretion, including the hormone gastrin and food. The H_2-receptor antagonist drugs, such as cimetidine and ranitidine, rapidly established a place in the treatment of conditions involving the hypersecretion of gastric acid, and they revolutionized the treatment of duodenal ulcers and the Zollinger-Ellison syndrome. Their safety record has been excellent, and they provide a valuable alternative to surgery.

Histamine's role in the stomach

Although the H_2-receptor antagonists drugs also block the effects of histamine on H_2 receptors at other sites (notably in the heart and blood vessels and on parts of the immune system), this has not proved to be a problem during therapeutic use. Histamine is not normally involved in other homeostatic mechanisms and even if it is, it affects systems that are also subject to control by many other messenger substances. There are circumstances in which it may be advantageous to administer H_1- and H_2-receptor antagonist drugs as the same time; *e.g.*, in treating the fall in blood pressure in anaphylactic shock (an exaggerated allergic reaction), involving both types of receptors.

Histamine has other actions in which its role is less well understood. It can stimulate the heart to beat faster or to increase its force of contraction. It can modify the responsiveness of various types of lymphocytes in the blood during the course of immunological reactions, and it may also affect the movements of specialized cells in the blood. Histamine storage sites and receptors, as well as the biochemical enzymes required for producing and disposing of histamine, are present in the brain; it is probable that histamine has a neurotransmitter role, but its function is unclear. The role of the histamine H_1 receptor is relatively unexplored. It can apparently regulate release of histamine in the brain and may therefore be involved in the transmission of neuronal signals.

Role of histamine as a drug

Histamine has had limited use as a drug. Applied topically, for example, in creams, it has been used to treat muscular pain, such as lumbago and fibrositis, and as a peripheral vasodilator to treat recurrent itching and swelling of toes, fingers, or ears (chilblains). A synthetic analogue that possesses some selectivity toward H_1 receptors is betahistine, which has had some use in preventing Menière's disease (vertigo, deafness, and other vestibular disturbances).

Histamine has also been used as a diagnostic agent. In the diagnosis of pheochromocytoma (a tumour of the adrenal medulla), histamine acts via an H_1 receptor to cause the liberation of norepinephrine and epinephrine from the tumour; this leads to a rise in blood pressure, whereas in a normal subject the blood pressure falls. Histamine acting via its H_2 receptor has been used to diagnose impairment of the acid-producing cells of the stomach; the absence of acid after an injection of histamine indicates that the acid-secreting glands are not functioning, a situation that results in pernicious anemia. A chemical analogue of histamine, betazole, is less potent but more selective as a stimulant of acid secretion. Impromidine is a potent synthetic analogue that is a highly selective stimulant for H_2 receptors.

Released histamine does not last long in the body; it is rapidly inactivated and disappears from the bloodstream within minutes. Its biologic inactivation occurs by one of

two mechanisms: (1) its side chain may lose its amine group (deamination) through oxidation by the enzyme diamine oxidase (histaminase); or (2) it may gain a methyl group (methylation) in the imidazole ring by the enzyme histamine-N-methyltransferase. (C.R.G.)

CHEMOTHERAPY

Chemotherapy, in its broadest sense, refers to the treatment of disease with chemicals. In current usage, however, the term applies primarily to the treatment of infectious diseases and cancer. When used to refer to infectious diseases, the term is antimicrobial chemotherapy. Antimicrobial chemotherapy can be used either for prophylaxis (prevention) or treatment (cure) of disease caused by bacteria, fungi, viruses, protozoa, or helminths. Cancer chemotherapy uses synthetic chemicals and antibiotics that can differentiate to some degree between normal tissue cells and cancer cells. Chemotherapy is used in the treatment of cancer; no therapeutic agents are available for the prevention of cancer. The dramatic progress made in the transplantation of tissue and organs has been due, in part, to the use of chemicals that modify the immune response in recipients of these tissue and organs.

The production and use of penicillin in the early 1940s became the basis for the era of modern antimicrobial chemotherapy. Streptomycin was discovered in 1944, and since then many other antibiotics have been found and put into use.

Modification of the basic structure of drugs

A major discovery following the introduction of antimicrobials to medicine was the finding that their basic structure could be modified chemically to improve their characteristics. The finding of a bacterium that produces the basic structural component responsible for the antimicrobial activity of penicillins and cephalosporins now permits engineering of compounds with vastly improved pharmacokinetics and with activity specific for certain microorganisms.

The discovery and use of antimicrobial agents has not eliminated infectious diseases, but it has dramatically reduced the predominance of many. The use of antimicrobial agents, however, has itself created problems. Some microorganisms have become resistant to drugs, requiring a continuing search for different (and often more expensive) agents. This increase in resistance to drugs has resulted from their widespread and sometimes indiscriminate use. In other cases, the use of antimicrobial agents has caused a change in the microbial ecology of humans and the environment. These resistant bacteria are commonly found in hospitals (nosocomial infections), where they can infect persons whose resistance already is decreased.

Basic concepts. An ideal chemotherapeutic agent is one that is cidal (kills) rather than static (inhibiting growth). It should affect a specific microbe or tissue cell and not affect other microbes or normal cells. It should be one to which the infectious organism or cancer does not become resistant and one that is not allergic or toxic to the host. An ideal chemotherapeutic agent must have pharmacological attributes favourable for its use. Therefore, if it is to affect organisms in the gastrointestinal tract, it must remain in the intestinal tract and not be absorbed or inactivated when given orally. If an oral drug is used to affect organisms in the blood or tissues, then it must be absorbed from the intestinal tract or it must be capable of being given parenterally (by injection). It must be able to penetrate tissues and be maintained for adequate periods of time at the site of the infection or cancer in concentrations sufficient to affect the microorganism or cancer cells.

The factors that affect these conditions are molecular size, ionic charge, rate of metabolism and excretion (half-life), lipid solubility, degree of protein binding, and presence or absence of inflammation in the tissue. The chemotherapeutic agent must be selectively toxic; that is, it should be active against the microorganism or cancer cell but it must not be toxic for the host in the amount administered.

Chemotherapeutic index

None of the chemotherapeutic agents presently in use meets all of these criteria. In fact, a number of compounds that produce significant toxic effects in humans are used because they have a favourable chemotherapeutic index; that is, the amount required for a therapeutic effect is below the amount that causes a toxic effect. The levels of these drugs in the patient must be carefully controlled so as not to exceed toxic levels.

Persons with certain altered organ functions, such as occurs in liver or kidney disease, are often especially susceptible to drug toxicity. Chemotherapeutic agents, however, can be used safely if drug concentrations in blood are measured, the dose adjusted to avoid toxic levels, and organ function or toxicity monitored closely.

Chemotherapeutic agents that are used in the treatment of disease are of three sources: (1) the synthetic chemicals; (2) chemical substances or metabolic products made by microorganisms, a group containing the antibiotics; and (3) plants.

Adverse effects. All chemotherapeutic agents can have adverse effects ranging from relatively harmless to serious and life-threatening, sometimes culminating in death. These effects can be due to direct toxicity; allergic or hypersensitivity reactions; or alterations in the numbers and types of micoorganisms that are the normal flora found in the mouth, intestine, vagina, skin, etc.

Direct toxicities are expressed in a variety of ways, and many of these are associated with the gastrointestinal tract (nausea, vomiting, and diarrhea) and skin rashes. They are usually minor and do not limit the use of the agent. In more extreme cases, the toxicities can result in serious damage to organs such as the kidneys, liver, and eyes and to the nervous system. Some chemotherapeutic agents affect normal red blood cells, which can result in anemia. Allergic or hypersensitivity reactions can range from minor effects such as skin rash and itching to more serious effects that include choking and difficulty in breathing. In some cases, a sudden and severe form of allergy (anaphylaxis) can result in death.

The use of antimicrobial agents, in particular the broad-spectrum agents (see below *Antibacterial drugs*), can result in an alteration in the number and type of microorganisms normally found on the skin and mucosal surfaces. This is due to the inhibitory activity of the antimicrobial agent on sensitive microorganisms found on these tissues. The eradication of some organisms relieves the inhibitory activity they have on each other, thereby allowing the surviving (resistant) organisms to multiply. In some cases, organisms such as yeast that are generally resistant to antibiotics increase to numbers sufficient to invade and infect tissue.

Categorization of antibacterial agents

Antibacterial drugs. Antibacterial agents are categorized as narrow-, broad-, or extended-spectrum agents. Narrow-spectrum agents (*e.g.*, penicillin G) affect primarily gram-positive bacteria. Broad-spectrum antibiotics, such as tetracyclines and chloramphenicol, affect both gram-positive and some gram-negative bacteria. An extended-spectrum antibiotic is one that, as a result of chemical modification, affects additional types of bacteria, usually gram-negative bacteria.

Whether an antimicrobial agent affects a microorganism depends on several factors. The drug must be delivered to a sensitive site in the cell, such as an enzyme that is involved in the synthesis of a cell wall or a protein or enzyme responsible for the synthesis of proteins, nucleic acids, or the cell membrane. Whether the antibiotic enters the cell depends on the ability of the drug to penetrate the outer membrane of the cell, or on the presence or absence of transport systems for the antimicrobial agent, or on the availability of channels in the cell surface. In some cases the microorganism prevents the entry of the antibiotic by producing an enzyme that destroys or modifies the antibiotic by transferring a chemical group. If the antimicrobial agent does not penetrate the organism or is destroyed or modified, or if the organism does not contain a sensitive site, then the microorganism will not be affected; in such a case it is said to be resistant.

A major problem associated with the use of antibacterial drugs is that an organism that originally was sensitive to a given drug can become resistant. For example, bacteria undergo spontaneous mutations; and exposure of these bacteria to an antimicrobial can eradicate sensitive organisms, thereby selecting a population resistant to that drug

and sometimes to related drugs. Bacteria sensitive to antimicrobial agents can become resistant by acquiring from resistant organisms deoxyribonucleic acid (DNA) containing genes coding for resistance (resistance genes). Bacteria sensitive to an antimicrobial can mate (conjugation) with bacteria containing resistance genes, or they can acquire these resistance genes by transduction. In transduction, a bacterial virus (bacteriophage) incorporates resistance genes into its genome by infecting a resistant bacterium. When the bacterial virus infects another bacterium, the phage DNA (containing resistance genes) can be incorporated into that bacterium and confer resistance. Some bacteria may acquire multiple resistance genes simultaneously and become resistant to several antibiotics. This is possible because circular pieces of DNA (plasmids) can, by recombination, acquire several genes, each of which codes for resistance to a different agent. Plasmids containing these multiple resistance genes can transfer to sensitive bacteria and thereby confer multiple resistance. Transfer of genes into the chromosome or into plasmids is facilitated in many cases because the genes are found on transposons, which are sequences of DNA that can excise themselves from plasmids and chromosomes and insert themselves into other plasmids and chromosomes. Bacteria resistant to as many as 10 different antimicrobial agents are known. One of the major problems associated with the transfer of resistance genes is that they can be transferred not only among similar but also to quite different bacteria.

Transposons

Resistance to antimicrobial agents results from (1) decreased permeability of the organism to the drug; (2) deactivation or modification of the drug by an enzyme; (3) modification of the drug receptor or binding site; (4) increased synthesis of an essential metabolite whose production is blocked by the antimicrobial agent; or (5) production of an enzyme that is altered so that it is not inhibited or affected by the drug.

Antibiotics. Antibiotics are substances produced by microorganisms that at low concentrations kill or inhibit other microorganisms. They are produced commonly by soil microorganisms and probably represent a means by which organisms in a complex environment, such as soil, control the growth of competing microorganisms. The microorganisms that produce antibiotics useful in preventing or treating disease include bacteria (*Bacillus* and *Streptomyces*) and fungi (*Penicillium, Cephalosporium,* and *Micromonospora*). Antibiotics can inhibit microbes by inhibiting the synthesis of the cell wall.

Other antibiotics, such as the aminoglycosides, chloramphenicol, erythromycins, and clindamycin, inhibit protein synthesis in bacteria. The basic process by which bacteria and animal cells synthesize proteins is similar, but the proteins involved are different. Those antibiotics that are useful as antibacterial agents (selectively toxic) utilize these differences to bind to or inhibit the function of the proteins of the bacterium, thereby preventing the synthesis of new proteins and new bacterial cells. Antibiotics such as polymyxin B and colistin bind to phospholipids in the cell membrane of the bacterium and interfere with its function as a selective barrier; this allows essential macromolecules in the cell to leak out, resulting in the death of the cell. Because other cells, including human cells, have similar or identical phospholipids, these antibiotics are somewhat toxic. One antibiotic, rifampin, interferes with RNA synthesis in bacteria by binding to a subunit on the bacterial enzyme responsible for duplication of RNA. Since the affinity of rifampin is much stronger for the bacterial enzyme than for the mammalian enzyme, the mammalian cells are unaffected at therapeutic dosages.

Bacteria, unlike animal cells, have a cell wall surrounding a cytoplasmic membrane. Production of the cell wall involves the partial assembly of wall components inside the cell, transport of these structures through the cell membrane to the growing wall, assembly into the wall, and finally cross-linking of the strands of wall material. Antibiotics that inhibit the synthesis of a cell wall have a specific effect on one or another phase. The result is an alteration in the cell wall and in the shape of the organism and the eventual death of the bacterium.

The penicillins and cephalosporins both have a unique structure, a β-lactam ring, that is responsible for their antibacterial activity. The β-lactam ring interacts with proteins in the cell responsible for the final step in the assembly of the cell wall. Thus, the mechanism of action is identical for both antibiotics; however, the basic chemical structure of the penicillins and cephalosporins differs in other respects, resulting in some difference in pharmacokinetics and the spectrum of antimicrobial activity.

The penicillins can be divided into two groups: the naturally occurring penicillins (penicillin G, penicillin V, and benzathine penicillin) and the semisynthetic penicillins. The semisynthetic penicillins are produced by growing the mold *Penicillium* under conditions whereby only the basic molecule (6-aminopenicillanic acid) is produced. By adding certain chemical groups to this molecule, several different semisynthetic penicillins are produced that vary in resistance to degradation by β-lactamase (penicillinase), an enzyme that specifically breaks the β-lactam ring, thereby inactivating the antibiotic. In addition, the antimicrobial spectrum of activity and pharmacological properties of the natural penicillins can be changed and improved by these chemical modifications.

The naturally occurring penicillins are important chemotherapeutic agents. Even after 40 years of use they are still the drugs of choice for treating streptococcal sore throat, tonsillitis, pneumococcal pneumonia, endocarditis caused by some streptococci, syphilis, gonorrhea, meningococcal infections, and infections caused by some anaerobic organisms. Several microorganisms, most notably the staphylococci, developed resistance to the naturally occurring penicillins, which led to the production of the penicillinase-resistant penicillins (methicillin, oxacillin, nafcillin, cloxacillin, and dicloxacillin).

To extend the usefulness of the penicillins to the treatment of infections caused by gram-negative rods, the broad-spectrum penicillins (ampicillin, amoxicillin, carbenicillin, and ticarcillin) were developed. These penicillins are sensitive to penicillinase, but they are useful in treating urinary tract infections caused by gram-negative rods as well as in treating typhoid and enteric fevers.

The extended-spectrum agents (mezlocillin, azlocillin, and piperacillin) are unique in that they have greater activity against gram-negative bacteria, including *Pseudomonas aeruginosa.* They have decreased activity, however, against penicillinase-resistant *Staphylococcus aureus.*

The penicillins are the safest of all antibiotics. The major adverse reaction associated with their use is hypersensitivity, with reactions ranging from a rash to bronchospasm and anaphylaxis. The more serious reactions are uncommon.

The cephalosporins are produced by *Cephalosporium acremonium.* Modification of the basic molecule (7-aminocephalosporanic acid) has resulted in three generations of cephalosporins. The first-generation cephalosporins (cefazolin, cephalothin, and cephapirin) have a range of antimicrobial activity similar to the broad-spectrum penicillins. The second-generation cephalosporins (cefamandole, cefonicid, cefotetan, cefoxitin, and cefuroxime) have greater β-lactamase stability than the earlier cephalosporins, and their antibacterial spectrum has been extended to include greater activity against additional species of gram-negative rods. They have decreased activity, however, against gram-positive bacteria. Like the penicillins, the cephalosporins are relatively nontoxic. Because the structure of the cephalosporins is similar to that of penicillin, hypersensitivity reactions can occur in penicillin-hypersensitive patients.

Generations of cephalosporins

Cycloserine, an antibiotic produced by *Streptomyces orchidaceus,* is a structural analogue of the amino acid D-alanine, and it interferes with enzymes necessary for incorporation of D-alanine into the bacterial cell wall. It is rapidly absorbed from the gastrointestinal tract and penetrates most tissues quite well; high levels are found in urine. It is used in the treatment of tuberculosis and in some urinary tract infections.

Bacitracin is produced by a special strain of *Bacillus subtilis.* Because of its toxicity its use is limited to the topical treatment of skin infections caused by streptococci and staphylococci and for eye and ear infections. Van-

comycin, an antibiotic produced by *Streptomyces orientalis*, is poorly absorbed from the gastrointestinal tract and is usually given by intravenous injection. It is an excellent antibiotic for the treatment of serious staphylococcal infections caused by strains resistant to the various penicillins.

The aminoglycosides (streptomycin; neomycin; paromomycin; kanamycin and its derivative, amikacin; tobramycin; netilmicin; and spectinomycin) are produced by *Streptomyces* species. Gentamicin is produced by the molds *Micromonospora purpurea* and *M. echinospora*. All of the aminoglycosides inhibit protein synthesis, although spectinomycin, which has a different structure, does so by a mechanism different from the other aminoglycosides. The aminoglycosides are poorly absorbed from the gastrointestinal tract, so, with some exceptions, they are given by intramuscular injection. Neomycin is toxic and is used topically. Because it is poorly absorbed from the gastrointestinal tract, paromomycin is used in the treatment of protozoal infections of the intestinal tract.

Streptomycin was the first of the aminoglycosides to be discovered and the second antibiotic used in chemotherapy. One of its more important uses had been as part of the combined therapy for tuberculosis. It still has some use in combination with penicillin for treating infections of heart valves (endocarditis) and with tetracyclines in the treatment of plague, tularemia, and brucellosis.

Kanamycin is used in the treatment of septicemia (blood poisoning), meningitis, and urinary tract infections caused by gram-negative bacteria. Because many organisms are resistant to its effects, however, kanamycin is now being replaced by other drugs. Gentamicin, tobramycin, netilmicin, and amikacin are similar in their range of antimicrobial activity. They are effective against infections caused by staphylococci and gram-negative bacteria, including *Pseudomonas aeruginosa*.

Margin of safety of aminoglycosides

The major problem with the aminoglycosides is that the margin of safety between a toxic and a therapeutic dose is narrow. Nephrotoxicity (harmful to kidney cells) and ototoxicity (harmful to the eighth cranial nerve of the organs of hearing and balance) are frequent, and the risk of these reactions increases with age and with preexisting renal diseases or hearing loss. Spectinomycin does not have the serious toxicity associated with the other aminoglycosides. It is used solely in treating gonorrhea in persons who are hypersensitive to penicillin or in those with gonococcal organisms resistant to penicillin.

Tetracyclines have a common structure but differ from each other by the presence or absence of chloride, methyl, and hydroxyl groups. Although these modifications do not change their broad-spectrum antimicrobial activity, they do affect pharmacological properties such as half-life in serum and protein-binding ability in serum. The tetracyclines all have the same antimicrobial spectrum, although there are some differences in sensitivity of the microorganisms to the various types of tetracyclines. They inhibit protein synthesis in both bacterial and animal cells. Bacteria have a system that allows tetracyclines to be transported into the cell, whereas animal cells do not; animal cells therefore are spared the effects of tetracycline on protein synthesis.

All tetracyclines are absorbed from the gastrointestinal tract after oral administration, and most can be given intravenously or intramuscularly. Because calcium, magnesium, aluminum, and iron form insoluble products with most tetracyclines, they cannot be given simultaneously with substances containing these minerals (*e.g.*, milk). They are the drugs of choice in the treatment of cholera, rickettsial infections, relapsing fever, trachoma (a chronic infection involving the eye), psittacosis (a disease transmitted by certain birds), brucellosis, tularemia, and respiratory infections. Tetracyclines are also used for acne vulgaris. Because not all of the orally administered tetracycline is absorbed from the gastrointestinal tract, the bacterial population of the intestine can become resistant to tetracyclines, resulting in overgrowth (suprainfection) of resistant organisms. Complexes between tetracyclines and calcium can cause staining of teeth and retardation of bone growth in young children or in the newborn if tetracyclines are taken by the mother after the fourth month of pregnancy. Tetracycline can also cause photosensitivity in patients exposed to sunlight.

Chloramphenicol now is synthesized chemically, but originally it was isolated from cultures of the bacterium *Streptomyces venezuelae*. It is administered either orally or parenterally, but since it is readily absorbed from the gastrointestinal tract, parenteral administration is reserved for serious infections. It is a broad-spectrum antibiotic used in the treatment of typhoid fever and for infections caused by microorganisms resistant to penicillin. Because newborns, particularly the premature newborn, cannot metabolize chloramphenicol, high levels accumulate and can cause inadequate oxygenation, the "gray syndrome." The most serious adverse effect is a toxic decrease in bone-marrow activity and aplastic anemia.

Erythromycin is produced by *Streptomyces erythreus*. It is usually administered orally, but it can be given parenterally. Although erythromycin has relatively few primary uses, it is valuable in treating pharyngitis and pneumonia caused by streptococci in persons sensitive to penicillin. It is also used in treating pneumonias caused either by *Mycoplasma* species or by the organism causing Legionnaire's disease; and it is used in treating pharyngeal carriers of the bacillus responsible for diphtheria.

Therapeutic indications of erythromycin

Clindamycin is a derivative of lincomycin that has better microbial activity and rate of gastrointestinal absorption. As a result, lincomycin has limited use. Clindamycin is active against staphylococci, some streptococci, and anaerobic bacteria. Because it has been associated with pseudomembranous colitis (inflammation of the small intestine and the colon), it must be used with caution. Other antibiotics, however, can cause an identical colitis.

The polymyxins are produced by *Bacillus polymyxa* and are designated as polymyxin A through E. Two of these, polymyxin B and polymyxin E (colistin), are useful in treating infection. Polymyxins B and E are polypeptide antibiotics with an affinity for phospholipids (important elements in cell membranes). Polymyxins accumulate in the cell membrane of bacteria and affect selective permeability. They also react with and affect the membranes of animal cells, resulting in kidney damage and neurotoxicity. Because they are not well absorbed from the gastrointestinal tract, oral administration is occasionally used for the treatment of diarrhea. Polymyxins can be administered by intramuscular injection. They are used primarily in treating infections caused by *Pseudomonas aeruginosa*, but they are also used topically for the treatment of eye and ear infections. The availability of other excellent antibiotics limits their use.

Rifampin, a semisynthetic agent derived from a rifamycin produced by *Streptomyces mediterranei*, inhibits RNA synthesis. It is absorbed from the gastrointestinal tract, penetrates tissue well, including the lung, and is used in the treatment of tuberculosis. Rifampin administration is associated with several side effects, mostly gastrointestinal in nature. The urine, feces, saliva, sweat, and tears can be red-orange in colour.

RNA inhibition

Synthetic chemical agents. Synthetic chemical agents used in treating bacterial diseases can affect bacteria in several ways. Some, such as the sulfonamides, are competitive inhibitors of essential biosynthetic pathways. They prevent the synthesis of folic acid, which is an essential preliminary step in the synthesis of nucleic acids. Sulfonamides are able to inhibit folic acid synthesis because they are similar to an intermediate compound (*p*-aminobenzoic acid) that is converted by an enzyme to folic acid. The similarity in structure between these compounds results in competition between *p*-aminobenzoic acid and the sulfonamide for the enzyme responsible for converting the intermediate to folic acid. This reaction is reversible by removing the chemical and results in the inhibition but not the death of the microorganisms.

Other synthetic chemical agents, such as isoniazid (INH) are similar to other metabolites (vitamins) and interfere with their utilization in the synthesis of essential cellular components. They may also activate an enzyme that destroys a factor essential to the cell. Methenamine, in an acidic environment (urine), decomposes to formaldehyde, which is antibacterial. Nalidixic acid affects the bacterial

cell by reacting with an enzyme that is required for the coiling of DNA. The nitrofurans are unique in that their reduction by the bacterium results in forms of oxygen that are highly reactive and thereby react with and alter essential cellular structures.

The sulfonamides are broad-spectrum agents and were once used widely. Their use has diminished because of the availability of antibiotics that are better and safer and because of increased instances of drug resistance. Sulfonamides are still used, but largely for treating urinary tract infections and preventing infection of burns. They are also used in the treatment of certain forms of malaria. The several forms (congeners) of sulfonamides differ from each other in solubility, half-life, ability to bind to plasma proteins, and potency for inhibiting certain bacteria. All affect bacterial growth by interfering with the synthesis of folic acid. Humans are not usually affected by the drugs because they do not synthesize folic acid but rather obtain it from their diet. Trimethoprim also affects the pathway of folic acid synthesis, but at a point different from that inhibited by the sulfonamides. When they are given together, the sequential blockage of the pathway produced by sulfamethoxazole and trimethoprim markedly enhances the inhibition of folic acid synthesis that would have been achieved by the activity of the sulfonamides alone. As a result, this combination is valuable in treating urinary tract infections, systemic infections, and intestinal tract infections caused by *Salmonella* or *Shigella* organisms. The sulfonamides are relatively safe, but hypersensitivity reactions (rashes, eosinophilia, and drug fever) can occur. Similar reactions are obtained with trimethoprim and with the sulfamethoxazole-trimethoprim combination.

Limited use of sulfonamides

The sulfones are related to the sulfonamides and are inhibitors of folic acid synthesis. They tend to accumulate in skin and inflamed tissue and are retained in the tissue for long periods. Thus, they are useful in treatment of leprosy.

Nalidixic acid, an agent commonly used for the treatment of urinary tract infections, affects an enzyme, termed DNA gyrase, that is required for the supercoiling of bacterial DNA.

The nitrofurans (nitrofurantoin and nitrofurazone) are broad-spectrum agents that undergo chemical reduction, resulting in the production of superoxide and other toxic oxygen compounds. These compounds oxidize essential components of the cell and make them nonfunctional. Nitrofurantoin is given orally, and because it accumulates in urine it is used in the treatment of urinary tract infections. Nitrofurazone is used topically for the treatment of burns.

Isoniazid, ethambutol, pyrazinamide, and ethionamide are synthetic chemicals used in treating tuberculosis. Isoniazid, ethionamide, and pyrazinamide are similar in structure to nicotinamide adenine dinucleoside (NAD), a coenzyme essential for several physiological processes. Ethambutol prevents the synthesis of mycolic acid, a lipid found in the tubercule bacillus. All of these drugs are absorbed from the gastrointestinal tract and penetrate tissues and cells. They are useful in treating tuberculosis. An isoniazid-induced hepatitis can occur, particularly in patients 35 years of age or older.

Antifungal drugs. The fungi appear in two morphological forms: a single cell that is round or oval (yeast), and a filamentous form (mold). Fungi differ from bacteria in several ways, including the chemical composition of the cell wall and cell membrane. Unlike bacteria, fungi have a nucleus surrounded by a membrane, an endoplasmic reticulum, and mitochondria. These differences between the bacteria and fungi are reflected in the use of different chemotherapeutic agents. Both antibiotics and chemical agents are used in the chemotherapy of fungal diseases.

Forms of fungi

Antibiotics. Amphotericin B and nystatin are antibiotics that interact with ergosterol, a type of steroid that is found in fungal membranes; this binding results in the loss of membrane-selective permeability and of cytoplasmic components. These antibiotics do not affect bacteria, because, with the exception of *Mycoplasma* species, bacteria do not have these types of steroids in the cell membrane. Mammalian cell membranes do, however, and there is some toxicity associated with parenteral use of these drugs. Amphotericin B is produced by *Streptomyces nodosus* and is used primarily in the treatment of serious fungal diseases, such as cryptococcal meningitis, histoplasmosis, and blastomycosis. The most serious side effect associated with amphotericin B is nephrotoxicity, although phlebitis and anemia can also result. Nystatin is produced by *Streptomyces noursei* and is used orally or topically for the treatment of mucocutaneous infections caused by *Candida albicans.* Nystatin is virtually nontoxic.

Griseofulvin, an antibiotic produced by *Penicillium griseofulvin,* is given orally for the treatment of several superficial fungal infections of the skin (*e.g.,* ringworm and athlete's foot) and diseases of the hair and nails. Griseofulvin binds to keratin, thus depositing high levels in the skin, which is the site of the infection. Griseofulvin affects the fungus by binding to microtubules, structures responsible for forming mitotic spindles and for processing cell wall components needed for growth.

Synthetic chemical agents. Several nonantibiotic preparations are used in treating fungal infections. Flucytosine (5-FC), a fluorinated pyrimidine, is unique in that it becomes active only when converted to 5-fluorouracil (5-FU) by an enzyme, cytosine deaminase, found in fungi but not present in mammalian cells. Flucytosine inhibits RNA and DNA synthesis. When administered parenterally, 5-FC is used primarily in the treatment of systemic cryptococcal and *Candida* infections and chronomycosis.

A group of antifungal agents, termed imidazoles, bind to fungal membranes and block synthesis of fungal lipids, especially ergosterol, the sterol required for membrane stability. The imidazoles have broad antifungal activity and are active against fungi that infect skin and mucous membranes and those that cause deep tissue infections. Clotrimazole, econazole, and tioconazole are given topically and are used for treating oral, skin, and vaginal infections. Miconazole can be administered topically, intravenously, or by intrathecal injection. It is used in treating skin infections, vaginitis, and systemic infections. Ketoconazole is readily absorbed after oral ingestion. It has broad-spectrum activity and is used in treating skin, mucocutaneous candidiasis, and several systemic fungal infections, such as blastomycosis.

Ciclopirox olamine, a hydroxypyridone, is a topical agent that accumulates in the skin and thus is effective in treating skin infections. It affects dermatophytic fungi (that infest the skin) by inhibiting the uptake of amino acids and other precursors needed for the synthesis of macromolecules. Tolnaftate and undecylenic acid are both used for treating dermatophyte infections.

Antiparasitic drugs. Although most organisms that live in or on humans are parasitic, the term parasite is commonly used in reference to the unicellular protozoans and the multicellular helminths (worms).

Antiprotozoal drugs. The protozoans, unlike bacteria and fungi, do not have a cell wall. They have a nucleus and a cytoplasm that is surrounded by a selectively permeable cell (plasma) membrane. The cytoplasm contains organelles similar to those found in other animal and plant cells (*e.g.,* mitochondria, Golgi apparatus, and endoplasmic reticulum). Thus, most of the antibiotics effective in inhibiting bacteria are not active against protozoans. Amphotericin B, however, reacts with sterols, which are components of both fungal and protozoal membranes. Most of the drugs used in the chemotherapy of diseases caused by the protozoans are derived from plants or are synthetic chemical compounds.

Structure of protozoans

Metronidazole is usually given orally for the treatment of vaginal infections caused by *Trichomonas vaginalis,* and it is effective in treating bacterial infections caused by anaerobes. It affects these organisms by causing nicks in, or breakage of, strands of DNA or by preventing DNA replication.

Iodoquinol inhibits several enzymes of protozoans. It is given orally for treating asymptomatic amebiasis and is given either by itself or in combination with metronidazole for intestinal and hepatic amebiasis. *Balantidium coli* and *Dientamoeba fragilis* infections also are treated with iodoquinol. Emetine, an alkaloid derived from ipecac syrup, is obtained from the roots of *Cephaelis ipecacuanha,* a plant native to Brazil. It is used along with iodoquinol

or chloroquine phosphate as alternative therapy for treatment of severe intestinal and hepatic amebiasis. Emetine is given by injection and can cause serious toxicity. Dehydroemetine is less toxic than emetine and may be used as an alternative drug.

Quinacrine is the drug of choice for giardiasis, an infection of the intestine caused by a flagellated amoeba. Quinacrine inserts itself into DNA, thereby ultimately preventing the synthesis of nucleic acids. It is given orally and can cause yellow staining of skin and sclera and deposition of blue and black pigment in the nail beds.

Trypanosomes are flagellated protozoans that cause a number of diseases. *Trypanosoma cruzi,* the agent of Chagas' disease, is treated with nifurtimox, a nitrofuran derivative. It is given orally and results in the production of activated forms of oxygen, which are lethal to the parasite. Other forms of trypanosomiasis (African trypanosomiasis, or sleeping sickness) are caused by *T. gambiense* or *T. rhodesiense.* When these parasites invade the blood or lymph, the drug of choice is suramin, a nonmetallic dye that affects glucose utilization and hence energy production. Because suramin is not absorbed from the gastrointestinal tract, it is given by intravenous injection. In the late form of trypanosomiasis, when the parasites have invaded the central nervous system, melarsoprol and tryparsamide are administered intravenously. They are used because they can penetrate the central nervous system and affect cellular structures and their functions.

Pneumocystis carinii causes pulmonary disease in immunocompromised patients. These infections are treated with trimethoprim-sulfamethoxazole, which inhibits folic acid synthesis in protozoans. An alternative agent for treatment of these diseases is pentamidine isethionate, which probably affects the parasite by binding to DNA. Because the drug is not well absorbed from the gastrointestinal tract, it is given by the intramuscular route.

Malaria

Malaria is one of the more serious protozoal infections. Chloroquine phosphate, given orally, is the drug of choice for prophylaxis and treatment. In regions where chloroquine-resistant *Plasmodium falciparum* is encountered, however, pyrimethamine, in combination with sulfadoxine, is used for prophylaxis. Both drugs interfere with folic acid synthesis and are well tolerated. Quinine sulfate, along with pyrimethamine and sulfadoxine, are used to treat infections caused by chloroquine-resistant *P. falciparum.* A high level of quinine in the plasma frequently is associated with cinchonism, a mild adverse reaction associated with such symptoms as a noise in the ears (tinnitus), headache, nausea, abdominal pain, and visual disturbance. Primaquine phosphate is given orally to prevent attacks after a person has left an area where *P. vivax* and *P. ovale* are endemic and to prevent relapses with the same organisms.

Anthelmintics. Helminths (worms) can be divided into three groups: cestodes, or tapeworms; nematodes, or roundworms; and trematodes, or flukes. The helminths differ from other infectious organisms in that they have a complex body structure. They are multicellular and have partial or complete organ systems (*e.g.,* muscular, nervous, digestive, and reproductive). Several of the drugs used to treat worm infections affect the nervous system of the parasite and result in muscle paralysis, either spastic or flaccid. Other drugs affect the uptake of glucose and thus energy stores. All are chemical agents and are generally administered orally. There are no antibiotics available for the treatment of these infestations.

Tapeworms attach to the intestinal tract by a sucker or a sucking groove on the head (scolex). Unlike the nematodes and trematodes, tapeworms do not enter the host tissues. The primary drugs used for these infections are niclosamide and praziquantel. Niclosamide inhibits the uptake of glucose by the helminth and therefore the production of energy. It has a spastic or paralytic effect on the worm. Because it is poorly soluble, high concentrations are obtained in the intestinal lumen. Praziquantel also produces tetanus-like contractions of the musculature of the worm. Unlike niclosamide, praziquantel is readily absorbed from the intestinal tract. It is a broad-spectrum anthelmintic affecting both flukes and tapeworms.

Various locations of roundworms

Treatment of roundworms is complicated by the fact that some live in blood, lymphatics, and other tissues (filarial worms) and thus require use of drugs that are absorbed from the intestinal tract and penetrate into tissues. Others are found primarily or solely in the intestinal tract (intestinal nematodes).

Diethylcarbamazine, used for treating filarial worm infections, is absorbed from the intestinal tract. Blood levels are reached quickly, and it has rapid action against the microfilariae. A severe allergic or febrile reaction due to the death of the microfilariae can follow use of the drug. Piperazine causes a flaccid paralysis of the worm and its expulsion from the intestinal tract. It is used as alternative therapy for treating *Ascaris* infection.

Thiabendazole and mebendazole interfere with glucose uptake and consequently with the production of energy. Mebendazole accumulates in the intestine and is used for treating *Ascaris,* hookworm, and whipworm infections. It is well tolerated but abdominal discomfort and diarrhea can occur in patients with a strong infestation. Thiabendazole is rapidly absorbed from the gastrointestinal tract, making it effective against organisms found in tissue. It is used in treating cutaneous larva migrans (creeping eruption), visceral larva migrans, trichinosis, and trichostrongylosis. About one-third of patients treated with thiabendazole have anorexia, nausea, vomiting, and vertigo.

Pyrantel pamoate causes spastic paralysis of helminth muscle. Most of the drug is not absorbed from the intestinal tract, resulting in high levels in the lumen. It is a drug of choice in treating pinworm and *Ascaris* infection and is a recommended alternative therapy for hookworm and trichostrongolosis.

Praziquantel is the most effective drug in treating infections caused by intestinal, liver, and lung flukes. Bithionol is used for treating *Fasciola hepatica* (sheep liver fluke) and *Paragonimus westermani* (lung fluke) infections and is absorbed from the intestinal tract. Tetrachloroethylene is an alternative agent for treating *Fasciolopsis buski* (large intestinal fluke) and *Metagonimus yokogawai* (small intestinal fluke) infection. Praziquantel is the drug of choice for treating schistosomiasis (infections of blood flukes). Metrifonate, a drug used as an alternative agent for *Schistosoma haematobium* infections, is metabolized to dichlorvos, an anticholinesterase agent. Dichlorvos acts on cholinesterase in the helminth and can also cause a reversible inhibition of plasma cholinesterase in patients. Oxamniquine in an alternative oral therapy for the treatment of *Schistosoma mansoni* infestation.

Antiviral drugs. Viruses are among the most common and widespread causes of infectious diseases. They cause such illnesses as influenza, herpes simplex type I (cold sores of the mouth) and type II (genital herpes), shingles, viral hepatitis, encephalitis, infectious mononucleosis, and the common cold. Viruses remain one of the least understood and most difficult of all infectious organisms to control; but this is changing as more is learned about their structure and replication. Viruses consist of nucleic acid, either DNA or RNA, and a protein coat. Because viruses do not have the enzymes that are needed to manufacture cellular components, they are obligate parasites, which means they must enter a cell for replication to occur.

Obligate parasites

The nucleic acid of the virus instructs the host cell to produce viral components, which leads to an infectious virus. In some cases, as in herpes infections, the virus nucleic acid may remain in the host cell without causing replication of the virus and damage to the host (viral latency). In other cases, the production of virus by the host cell may cause the death of the cell. A major problem in treating some viral diseases is that latent viruses can become activated, frequently when the host undergoes stress, thereby producing infectious virus and cellular effects.

Many factors account for the difficulty in developing antiviral chemotherapeutic agents. The structure of each virus differs, and specific therapy is often unsuccessful because of periodic changes in the antigenic proteins of the virus. The need for a host cell to support the multiplication of the virus makes treatment difficult because the chemotherapeutic agent must be able to inhibit the virus without seriously affecting the host's cells.

The greatest success against virus infections has been by increasing immunity through vaccination (influenza, poliomyelitis, measles, mumps, and smallpox) with live attenuated (weakened) or killed viruses. Vaccination has resulted in the total elimination of smallpox. While vaccination has proved to be effective against the specific virus used for smallpox, influenza is caused by viruses that constantly change their antigenic protein, thereby requiring revaccination as the antigenic makeup of the virus changes. Some virus groups contain 50 or more different viruses.

Passive immunization with serum or globulin (antibodies) from immune persons has been used to prevent viral infections. Immune globulins, such as those used against hepatitis, often cause adverse effects, however, and they are effective only for prophylaxis and not for treatment.

Viral replication cycle

An antiviral agent must act at one of five basic steps in the viral replication cycle in order to inhibit the virus. The steps are (1) attachment and penetration of the virus into the host cell; (2) uncoating of virus—*e.g.*, removal of the protein surface and release of the viral DNA or RNA; (3) synthesis of new viral components by the host cell as directed by the virus DNA; (4) assembly of the components into new virus; and (5) release of the virus from the host cell.

Antibiotics. Vidarabine (adenine arabinoside) was isolated from *Streptomyces antibioticus.* It is given intravenously and is effective in treating herpes simplex types I and II, particularly severe herpes encephalitis infections, herpes keratitis (eye infection), and varicella zoster (shingles) infections. Vidarabine is similar in structure to a component of viral DNA (a purine). The host cell adds chemical groups (phosphates) to the antibiotic, and this change causes inhibition of DNA polymerase, an enzyme that catalyzes the formation of DNA. It also incorporates into the DNA molecule of the virus and prevents formation of the DNA strand. Vidarabine does not affect host cells at the concentration used for the inhibition of viral replication. Mutants that code for altered DNA polymerase are not affected by this antibiotic.

Synthetic chemical agents. Amantadine is an oral drug used for the prevention and treatment of the influenza A virus; it is not effective, however, against influenza B virus. The action of amantadine is probably through the inhibition of viral uncoating, thereby preventing the release of viral nucleic acid in the host cell and the synthesis of new virus. Amantadine may also inhibit penetration of the virus into the host cell. A similar agent, rimantadine, has the same mechanism of action but has a lower incidence of adverse effects (nervousness, confusion, drowsiness, and headache) than amantadine.

Treatment of herpes simplex infections

Acyclovir, which can be given orally, topically, or intravenously, is useful in the prevention and treatment of infections with herpes simplex type I and type II, shingles, and herpes keratitis. Acyclovir prevents replication of viral DNA either by inhibiting viral DNA polymerase or by incorporation into the viral DNA. Acyclovir is phosphorylated within infected cells by the action of the enzyme thymidine kinase, which means that the virus is inhibited by the drug more effectively than is the host cell, thus accounting for acyclovir's selective toxicity. Virus mutants that code for an altered thymidine kinase with reduced affinity for the drug are resistant to the drug's effects.

Idoxuridine and trifluorothymidine are used topically to treat herpes simplex keratitis, and both serve as substrates for the herpesvirus-induced DNA polymerase and consequently inhibit production of viral DNA.

Interferons represent a group of nonspecific antiviral proteins produced by host cells in response to viral infections, as well as in response to the injection of double-stranded RNA, some protozoal and bacterial components, and other chemical substances. Interferon results in the production of a protein that prevents the synthesis of viral components from the viral nucleic acid template. The interferons are of interest because they have broad-spectrum antiviral activity and because they inhibit the growth of cancer tissue. The study of interferons has been hampered somewhat because only small amounts are produced by tissue cells. The gene for the production of interferon, however, has been inserted into bacteria by recombinant DNA techniques, thus allowing production of large amounts of antiviral substance, a requirement for further testing and general use. Experimental trials have indicated the potential usefulness of interferon in the treatment of viral infections and some forms of cancer.

CANCER CHEMOTHERAPY

Mechlorethamine, an alkylating agent, is a derivative of nitrogen mustard, a chemical warfare agent. Mechlorethamine was first used in the 1940s in the treatment of cancer and was shown to be effective in treating human lymphomas. Since then, many anticancer (antineoplastic) drugs have been developed and used with much success. At least 10 types of human cancer can be cured in most patients by chemotherapy alone or in combination with surgery and/or radiation.

Toxicity of anticancer agents

Unlike other chemotherapeutic agents, where the goal is to destroy a microbial invader, the treatment of cancer is complicated in that the chemotherapeutic agent is aimed toward human cells, albeit cells that have undergone genetic changes and are dividing at a fast and uncontrolled rate. Because cancer cells are similar to normal human cells, the anticancer agents are generally toxic to normal cells and can cause numerous side effects, some of which are life threatening. These side effects include hair loss, sores in the mouth and on other mucous membranes, cardiac anomalies, bone marrow toxicity, and severe nausea and vomiting. The bone marrow toxicities result in anemia as well as in decreased resistance to infectious agents. Permanent infertility can also result. These adverse effects may require that the drug dosage be reduced or the antineoplastic drug regimen be changed to make the drug tolerable to the patient. Many of the anticancer drugs that are given intravenously cause severe local damage at the injection site if inadvertently injected into the tissue instead of the vein. The person preparing and administering the drugs also must be careful not to come in direct contact with these drugs because some can cause cancer (carcinogenic). While most are administered intravenously, many can be taken orally, and some can be injected intramuscularly or intrathecally (within the spinal cord).

Antineoplastic agents are divided into categories based on their mode of action. Since most of the drugs exert their effects in a certain part of the cell cycle (cell growth phase, cell division phase, resting phase, etc.), many treatment regimens require two or more of these agents. One drug may be used to stop the growth of the cancer cells in a certain phase, whereas another agent may work at a different phase. Using multiple agents, therefore, lessens the incidence of cellular resistance to an antineoplastic agent. The use of multiple agents also often enables the use of lower dosages of each drug, thereby reducing the side effects caused by each. In addition to using complex regimens that employ several drugs, chemotherapy is often combined with surgery to reduce the number of cancer cells and with radiation treatment to further destroy the cells.

Mechanisms of action

Alkylating agents. Alkylating agents were the first anticancer drugs used, and despite their hazards they remain a cornerstone of anticancer therapy. Some examples of alkylating agents are nitrogen mustards (chlorambucil and cyclophosphamide); cisplatin; urea derivatives (carmustine, lomustine, and semustine); alkylsulfonates (bisulfan); ethyleneimines (thiotepa); and triazenes (dacarbazine). These chemical agents are highly reactive and bind to certain chemical groups (phosphate, amino, sulfhydryl, hydroxyl, and imidazole groups) commonly found in nucleic acids and other macromolecules. These agents bring about changes in the DNA and RNA of both cancerous and normal cells. For example, the nucleic acid may lose a basic component (purine), or it may break, or strands of DNA may cross-link. The result is that the nucleic acid will not be replicated. The altered DNA either will be unable to carry out the functions of the cell, resulting in cell death (cytoxicity), or the altered DNA will change the cell characteristics, resulting in an altered cell (mutagenic change). This change may result in the ability or tendency to produce cancerous cells (carcinogenicity). Normal cells

may also be affected and become cancer cells. The alkylating agents can cause severe nausea and vomiting as well as decreases in the number of red and white blood cells. The decrease in the number of white blood cells results in susceptibility to infection. Although alkylating agents may be used in most types of cancer, they are generally of greatest advantage in treating slow-growing cancers. They are not as effective on rapidly growing cells.

Antimetabolites. Antimetabolites are antineoplastic agents that are structurally similar to compounds that are found naturally in the host (vitamins, amino acids, or precursors of DNA and RNA). They incorporate into either DNA or RNA (purine and pyrimidine nucleotides) and interfere with cellular function. Some inhibit an enzyme necessary to macromolecular synthesis, thereby preventing synthesis of essential host materials. Examples of these include antagonists of purines (azathioprine, mercaptopurine, and thioguanine) and antagonists of pyrimidine (fluorouracil and floxuridine). Cytarabine, which also has antiviral properties, interferes with a DNA polymerase, dihydrofolate reductase, which is necessary for the synthesis of tetrahydrofolate and subsequently for the synthesis of the folic acid needed for DNA formation.

Because the antimetabolites primarily act upon cells undergoing synthesis of new DNA for formation of new cells, it follows that most of the toxicities associated with these drugs are seen in cells that are growing and dividing quickly. They are known to cause severe damage to the mucous membranes of the mouth and other parts of the gastrointestinal tract and also to produce skin disorders and hair loss. Anemia can occur, along with a decrease in the number of white blood cells, which are necessary to prevent infections. Methotrexate has been used in low doses for the treatment of rheumatoid arthritis.

Antineoplastic antibiotics. Antineoplastic antibiotics (doxorubicin, daunorubicin, bleomycin, mitomycin, and dactinomycin) are derived from *Streptomyces* species. While they may have antibacterial activity, they are generally too dangerous and toxic for that use. These antibiotics affect DNA synthesis and replication by inserting into DNA or by donating electrons which result in the production of highly reactive oxygen compounds (superoxide) that cause breakage of DNA strands. These agents are associated with blood cell damage, hair loss, and other toxicities common to the antimetabolites and alkylating agents, and severe cardiac or lung toxicity also results. The effects vary in proportion to the dose and length of treatment. These antibiotics are administered exclusively by intravenous infusion.

Hormones. Hormones are used primarily in the treatment of cancers of the breasts and the sex organs. These tissues require hormones, such as androgens, progestins, or estrogens, for growth and development. By countering these hormones with an antagonizing hormone, the growth of that tissue is inhibited, as is the cancer growing in the area. For example, estrogens are required for female breast development and growth. Tamoxifen competes with endogenous estrogens for receptor sites in breast tissue where the estrogens normally exert their actions. The result is a decrease in the growth to breast tissue and of breast cancer tissue. Adrenocorticosteroids are also used for treating some types of cancer. An unusual approach to cancer chemotherapy has been the use of a hybrid molecule (estramustine) that is a complex of an estrogen and a nitrogen mustard. The hormones are the closest approach to a site-specific antineoplastic drug, but they work only on certain types of cancer.

Antagonizing effects

Other agents. A number of other agents are used in the treatment of cancer. Vinblastine and vincristine (vinca alkaloids), derived from the periwinkle plant, along with etoposide, primarily act to stop spindle formation within the dividing cell during DNA replication and cell division. Etoposide, a semisynthetic derivative of a toxin found in roots of the American mandrake, or may-apple plants, affects an enzyme and causes breakage of DNA strands. Hydroxyurea inhibits the enzyme ribonucleotide reductase, an important element in DNA synthesis. It is used to reduce the high granulocyte count found in chronic myelocytic leukemia. Asparaginase breaks down the amino acid asparagine to aspartic acid and ammonia. Some cancer cells, particularly in certain forms of leukemia, require this amino acid for growth and development. Other agents, such as dacarbazine and procarbazine, act through various methods, although they can act as alkylating agents. Mitotane, a derivative of the insecticide DDT, causes necrosis of adrenal glands.

New approaches to cancer therapy. Other agents being evaluated include interferon and monoclonal antibodies. Interferon is an antiviral agent that is effective in some cancers. Monoclonal antibodies are highly specific agents that recognize differences between cells. The approach to their use is twofold: first, to use the antibody to kill cancer cells by direct action; and second, to couple a toxin to the antibody. The antibody is used to recognize a specific cancer cell and to deliver a toxin that kills it.

Interferon and monoclonal antibodies

The BCG vaccine is an attenuated form of the bacillus that causes tuberculosis. It is used to immunize selected persons so as to prevent tuberculosis, and it has been used in the treatment of certain forms of cancer. Other chemicals are used to enhance the immunity to cancer cells.

The decision to use a certain antineoplastic drug depends on many factors, including the type and location of the cancer, its severity, whether surgery or radiation therapy can or should be used, and the side effects associated with the drug. Combinations of anticancer drugs, like combinations of antimicrobials, are often more effective than single agents. Although many regimens have been developed, each must be tailored to the patient. (I.S.S.)

IMMUNOSUPPRESSANTS

The immunosuppressants are a class of drugs capable of inhibiting the body's immune system. Many of the agents included in this category are also cytotoxic (cell poisons) and are used in the treatment of cancer. Cytotoxic agents used as immunosuppressants include antimetabolites (*e.g.*, azathioprine), alkylating agents (*e.g.*, cyclophosphamide), and folic-acid antagonists (*e.g.*, methotrexate; see above *Cancer chemotherapy*). Other immunosuppressants include corticosteroids, prednisone, antilymphocyte serum (ALS), and cyclosporin A. Radiation (*e.g.*, X rays) is also used to suppress the body's immune system.

The action of most cytotoxic drugs or hormonal agents is nonspecific; they may also act upon components of the immune system that are beneficial. Cytotoxic agents are capable of killing immunologically competent cells, but they do not preferentially affect lymphocytes. The cytotoxic agents, with their inherent ability to kill any cell that will replicate, were originally synthesized as antineoplastic drugs. It was soon discovered, however, that these drugs not only inhibit tumour cell growth but are also able to suppress the cell of the immune system. Subsequently, these cytotoxic agents were employed clinically to suppress the immune system in patients undergoing organ (*e.g.*, kidney, liver, or heart) transplants. The therapeutic use of these cytotoxic agents is based on their ability to suppress the recipient's immune system in order to prevent organ or graft rejection. Although most of the cytotoxic agents exhibit immunosuppressant properties, cyclophosphamide and azathioprine are the agents that have been used most frequently in patients who are preparing to undergo organ transplant surgery.

Use in organ transplant surgery

Cyclophosphamide can impair humoral immune reactions as well as cell-mediated immune response. Despite its therapeutic usefulness, cyclophosphamide is an exceedingly toxic drug and has a number of undesirable side effects. Azathioprine exerts its pharmacological action by inhibiting several enzymatic pathways required for the synthesis of DNA. Azathioprine is more effective in suppressing proliferating (dividing) lymphocytes; it is less active against nondividing cells. Like cyclophosphamide, azathioprine is a relatively toxic drug.

Although corticosteroids have been the primary agents employed in immunosuppressant therapy, their precise mechanism of action remains uncertain. It is known that the corticosteroids affect leukocyte function, alter lymphocyte populations, and depress certain serum immunoglobulins. The use of these steroids may lead to undesirable side effects.

Cyclosporin A is a metabolite obtained from a fungus. It depresses certain subpopulations of white blood cells (T lymphocytes and, to some extent, B lymphocytes). Cyclosporin A shows promise in reducing allograft (a transplant between genetically different members of the same species; *i.e.,* persons who are not identical twins) rejection in surgical transplants and may also be of value in the treatment of certain autoimmune diseases. (J.A.T.)

BIBLIOGRAPHY

General works: VICTOR A. DRILL, *Drill's Pharmacology in Medicine,* 4th ed., edited by JOSEPH R. DIPALMA (1971); and J.H. GADDUM, *Gaddum's Pharmacology,* 9th ed., edited by A.S.V. BURGEN and J.F. MITCHELL (1985), both classic texts; ALFRED GOODMAN GILMAN *et al.* (eds.), *Goodman and Gilman's The Pharmacological Basis of Therapeutics,* 7th ed. (1985), a comprehensive source with an emphasis on the clinical applications of drugs; AVRAM GOLDSTEIN, LEWIS ARONOW, and SUMNER M. KALMAN, *Principles of Drug Action: The Basis of Pharmacology,* 2nd ed. (1973), broad and detailed coverage of the physiological effects of drugs; JACK R. COOPER, FLOYD E. BLOOM, and ROBERT H. ROTH, *The Biochemical Basis of Neuropharmacology,* 5th ed. (1986), an introduction to many aspects of neuropharmacology; W.C. BOWMAN and M.J. RAND, *Textbook of Pharmacology,* 2nd ed. (1980), a comprehensive treatment of the physiological and biochemical processes underlying pharmacological mechanisms; H.O. SCHILD, *Applied Pharmacology,* 12th ed. (1980), an introduction; CHARLES R. CRAIG and ROBERT E. STITZEL (eds.), *Modern Pharmacology,* 2nd ed. (1986); and ANDRES GOTH, *Medical Pharmacology: Principles and Concepts,* 11th ed. (1984). The basis of drug action is studied in WILLIAM C. HOLLAND, RICHARD L. KLEIN, and ARTHUR H. BRIGGS, *Introduction to Molecular Pharmacology* (1964); and E.J. ARIËNS (ed.), *Molecular Pharmacology: The Mode of Action of Biologically Active Compounds,* 2 vol. (1964). Mechanisms of drug action are examined in JOHN W. LAMBLE (ed.), *Towards Understanding Receptors* (1981), *More About Receptors* (1982), and JOHN W. LAMBLE and ALISON C. ABBOTT (eds.), *Receptors, Again* (1984), collections of articles from pharmacological journals. Effects of drugs on the body are studied in STEPHEN H. CURRY, *Drug Disposition and Pharmacokinetics,* 3rd ed. (1980); JOHN W. LAMBLE (ed.), *Drug Metabolism and Distribution* (1983); and MILO GIBALDI and LAURIE PRESCOTT (eds.), *Handbook of Clinical Pharmacokinetics* (1983).

Types of drugs: (*Autonomic nervous system pharmacology*): MICHAEL D. DAY, *Autonomic Pharmacology: Experimental and Clinical Aspects* (1979), a general text; and STANLEY KAUSNER (ed.), *Trends in Autonomic Pharmacology,* 2 vol. (1985–82), a collective study. (*Central nervous system pharmacology*): W.D. WYLIE, *Wylie and Churchill-Davidson's A Practice of Anaesthesia,* 5th ed., edited by H.C. CHURCHILL-DAVIDSON (1984), a comprehensive reference source; THOMAS E. KEYS, *The History of Surgical Anesthesia,* rev. ed. (1963, reprinted 1978), an authoritative account; and JOHN ADRIANI, *Labat's Regional Anesthesia: Techniques and Clinical Applications,* 4th ed. (1985), a comprehensive, well-illustrated text. Other studies of physiological mechanisms and applications of anesthetics include ROBERT D. DRIPPS, JAMES E. ECKENHOFF, and LEROY D. VANDAM, *Introduction to Anesthesia: The Principles of Safe Practice,* 6th ed. (1982); and RUDOLPH H. DE JONG, *Local Anesthetics,* 2nd ed. (1977). Analgesics and narcotics are treated in JOHN J. BONICA (ed.), *Pain* (1980), an examination of the body systems involved in pain and of the physiological, psychological, and clinical aspects of therapy. Drugs that affect mood and behaviour are the subject of JACK D. BARCHAS *et al.* (eds.), *Psychopharmacology: From Theory to Practice* (1977), an introductory text with detailed examples of treatment protocols and problems; MORRIS A. LIPTON, ALBERTO DIMASCIO, and KEITH F. KILLAM (eds.), *Psychopharmacology: A Generation of Progress* (1978), a general historical analysis; S.J. ENNA, JEFFREY B. MALICK, and ELLIOTT RICHELSON (eds.), *Antidepressants: Neurochemical, Behavioral, and Clinical Perspectives* (1981), a symposium of papers by leading practitioners; F. NEIL JOHNSON (ed.), *Handbook of Lithium Therapy* (1980), an analysis of the therapeutic use of lithium in the treatment of mental disorders; and JUDITH P. SWAZEY, *Chlorpromazine in Psychiatry: A Study of Therapeutic Innovation* (1974), a historical study. Sedatives are analyzed in D.J. GREENBLATT, R.I. SHADER, and D.R. ABERNETHY, "Drug Therapy: Current Status of Benzodiazepins," *The New England Journal of Medicine,* 309(6):354–358 (Aug. 11, 1983) and 309(7):410–416 (Aug. 18, 1983). Applications of antiepileptic drugs are the topic of GAIL E. SOLOMON, HENN KUTT, and FRED PLUM, *Clinical Management of Seizures: A Guide for the Physician,* 2nd ed. (1983); and H.-H. FREY and D. JANZ (eds.), *Antiepileptic Drugs* (1985).

(*Cardiovascular system pharmacology*): R. DOUGLAS WILKERSON (ed.), *Cardiac Pharmacology* (1981), an overview of the drug therapy and physiological effects of cardiovascular agents; K. GREEFF (ed.), *Cardiac Glycosides,* 2 vol. (1981), and PHILIP NEEDLEMAN (ed.), *Organic Nitrates* (1975), collections of review articles on the basics of cardiovascular pharmacology; E.M. VAUGHAN WILLIAMS, *Antiarrhythmic Action and the Puzzle of Perhexiline* (1980), an account of different types of antidysrhythmic drugs; PETER H. STONE and ELLIOTT M. ANTMAN (eds.), *Calcium Channel Blocking Agents in the Treatment of Cardiovascular Disorders* (1983), a collection of papers on the therapeutic uses of this group of drugs; and P.A. VAN ZWIETEN (ed.), *Pharmacology of Antihypertensive Drugs* (1984), a survey of hypotensive agents and their use in therapy.

(*Drugs affecting blood*): ROBERT W. COLMAN *et al.* (eds.) *Hemostasis and Thrombosis: Basic Principles and Clinical Practice* (1982), a comprehensive treatment of drugs used in blood coagulation disorders; H.E. KARGES and N. HEIMBURGER (eds.), *Aspects of Blood Coagulation and Fibrinolysis* (1983), a collection of research articles; DAVID BERGQUIST, *Postoperative Thromboembolism: Frequency, Etiology, Prophylaxis* (1983; originally published in Swedish, 1981), an analysis of the disorder and its therapy; and GESINA L. LONGENECKER (ed.), *The Platelets: Physiology and Pharmacology* (1985), a study of the effects of drugs on blood platelets. More advanced information on specific aspects of hemostasis and drug action is to be found in specialized journal articles such as W.H. FRISHMAN, "Antiplatelet Therapy in Coronary Heart Disease," *Hospital Practice,* 17(5):73–86 (May 1982); G.V.R.K. SHARMA *et al.,* "Thrombolytic Therapy," *The New England Journal of Medicine,* 306(21):1268–76 (May 27, 1982).

Specific groups of drugs are studied in the following works: W.C. BOWMAN, *Pharmacology of Neuromuscular Function with Special Reference to Anesthetic Practice* (1980), an examination of drugs affecting neuromuscular transmission; EDITH BÜLBRING (ed.), *Smooth Muscle: An Assessment of Current Knowledge* (1981), a collection of articles on anatomical, physiological, biochemical, and pharmacological aspects; SUSAN M. BARLOW and FRANK M. SULLIVAN, *Reproductive Hazards of Industrial Chemicals: An Evaluation of Animal and Human Data* (1982); JOHN A. THOMAS and EDWARD J. KEENAN, *Principles of Endocrine Pharmacology* (1986), an analysis of the effects of drugs on reproductive organs; and DONALD W. SELDIN and GERHARD GIEBISCH (eds.), *The Kidney: Physiology and Pathophysiology* (1985), a study of drugs used in treating kidney diseases. For information on topically applied drugs, the above mentioned general sources are useful, as are FREDERICK H. MEYERS, ERNEST JAWETZ, and ALAN GOLDFIEN, *Review of Medical Pharmacology,* 7th ed. (1980); and KENNETH L. MELMON and HOWARD F. MORRELLI (eds.), *Clinical Pharmacology: Basic Principles in Therapeutics,* 2nd ed. (1978). The comprehensive texts cited above also cover hormones and other body chemicals used as drugs; see also MAURICIO ROCHA E SILVA (ed.), *Histamine II and Anti-Histaminics: Chemistry, Metabolism, and Physiological and Pharmacological Actions* (1978); and C.R. GANELLIN and M.E. PARSONS (eds.), *Pharmacology of Histamine Receptors* (1982). For discussion of the many aspects of antimicrobial therapy and chemotherapy, see *Drug Facts and Comparisons* (annual, with monthly loose-leaf updates), which describes mechanisms of action, pharmacological properties, and recommended uses; MARK ABARAMOWICZ (ed.), *Handbook of Antimicrobial Therapy,* rev. ed. (1980), a reference source that gives a summary of antimicrobial agents, their use for specific diseases, doses, costs, and adverse effects; VICTOR LORIAN (ed.), *Antibiotics in Laboratory Medicine,* 2nd ed. (1986), a description of the methods used for the measurement of antimicrobial agents and their effects; K.G. NICHOLSON, "Antiviral Therapy: Respiratory Infections, Genital Herpes, and Herpetic Keratitis," *The Lancet,* 2(8403):617–621 (Sept. 15, 1984); RAPHAEL DOLIN, "Antiviral Chemotherapy and Chemoprophylaxis," *Science,* 227(4692):1296–1303 (March 15, 1985); and DANIEL P. STITES *et al.* (eds.), *Basic & Clinical Immunology,* 5th ed. (1984).

(H.P.R./C.R.G./J.S.F./A.W.C./J.A.T./I.S.S.)

Dublin

Dublin (Irish: Dubh Linn), the preeminent city of the republic of Ireland, is the political, commercial, and cultural capital of the nation. It is a city of contrasts, maintaining an uneasy relationship between reminders of its colonial past and symbols of present-day life.

This article is divided into the following sections:

Physical and human geography 561
- The landscape 561
  - Site
  - Climate
  - Layout
- The people 562
  - Demography
  - Religion
- The economy 562
  - Industry
  - Finance and commerce
  - Transportation
- Administration and social conditions 563
  - National and local government
  - Health
  - Education
- Cultural life 563
  - Theatre and music
  - Publishing
  - Sports

History 563
- Foundation and early growth 563
- Ascendancy in the 18th century 564
- Evolution of the modern city 564
- After national independence 564

Bibliography 565

Physical and human geography

THE LANDSCAPE

Site. Dublin's geographic site is superb. Situated at the head of a lovely bay, the city straddles the River Liffey where that stream flows eastward through a hill-ringed plain to the shores of the Irish Sea. (The dark bog water made the "black pool"—Dubh Linn in Irish, Dyfflin in Norse—that gave the city its name.) Almost certainly it was this opening from the sea, leading through the mountains to the fruitful central plains of Ireland, that originally tempted wandering Norse raiders to settle there. In spite of its long historical development, Dublin remains a physically small city. From Dublin Castle it is little more than four miles (six kilometres) to the farthest city boundary in any direction. Each year the suburbs jut farther into the countryside, but to the south there is a natural limit posed by the Dublin and Wicklow mountains, which ring the city and provide some of its most beautiful urban vistas.

Climate. With its coastal site on the western seaboard of the Irish Sea, Dublin enjoys a mild climate. The average temperature is lowest in January–February, 42° F (6° C), and highest in July–August, 59° F (15° C). Sunshine averages four hours a day. The mean annual rainfall is 30–40 inches (760–1,000 millimetres), although the rate is higher in the mountains. The period of maximum rainfall occurs in winter, and there are fewer than 10 days of snow a year.

Layout. Dublin is a low-built, steepled city, with few buildings dating from before the 17th century. The Roman Catholic churches are 19th- and 20th-century structures. One of the tallest buildings—Liberty Hall, a trade union headquarters—reaches 17 stories in height, but most of the buildings are no higher than 10 stories.

Norse, Norman, and Georgian, the three elements that constitute the architectural legacy of Dublin, all meet in Dublin Castle. In the first two decades of the 13th century, the Normans obliterated the Viking stronghold and reared a château-fort. When the Georgians built the present red-brick castle, they left two towers of the old structure standing. The castle, the seat of British authority in Ireland until 1922, is now used for ceremonial occasions, especially the inauguration of the republic's presidents, who now reside at Áras an Uachtaráin ("The President's House," formerly the viceroy's lodge) in Phoenix Park. Dublin Castle

Close to the castle a Viking king of Dublin built Christ Church Cathedral (*c.* 1030), which was replaced about 140 years later by a more magnificent Norman structure. By the 19th century the edifice was in ramshackle condition; it was restored in the 1870s at enormous cost. Its neighbour, St. Patrick's, erected just outside the city walls, was also originally a Viking church that may have been built on an earlier Celtic foundation. The Normans rebuilt it in 1191, and it was enlarged and partially rebuilt over the centuries. It was in a state of collapse when Sir Benjamin Lee Guinness, the brewing magnate and a lord mayor of Dublin, financed its restoration in the mid-19th century. Christ Church is the cathedral for the Protestant diocese of Dublin and Glendalough, whereas St. Patrick's, also Protestant, is the national cathedral. Both have been Church of Ireland (Anglican) churches since the Reformation.

The area between St. Patrick's and the Guinness Brewery on the Liffey is known as the Liberties, having formerly been outside the city walls and under the sole jurisdiction of the archbishop. Since World War II large tracts of this district have been cleared for low-cost housing.

Dublin's early private speculators had a sense of order and beauty as acute as their sense of profit. The city's streets were broad and its garden squares spacious. For their time (the 18th century), the houses were ultramodern—elegant yet simple Georgian and Neoclassical structures designed in the manner of the great English architects Inigo Jones and Sir Christopher Wren. The sweep of red-brick houses, ranged in long terraces, with their well-proportioned windows made a harmonious whole that still stands as a felicitous achievement of urban architecture.

In the southern half of the town, between Trinity College and St. Stephen's Green, Joshua Dawson, one of Dublin's leading citizens, built an impressive residence in 1705. A decade later he sold it to the city of Dublin for the lord mayor's residence, and it still serves this purpose. It was there that the first Irish republican parliament, the Dáil Éireann, met in 1919.

Dawson's neighbours, the equally prominent Molesworths, followed his example and began building houses and entire streets. In 1745–48 the Earl of Kildare erected, at the end of Molesworth Street, a palace, Kildare House, that was renamed Leinster House when he became duke of Leinster. Leinster House is now the seat of the Irish Parliament. Twin Victorian buildings, which were constructed on either side of Leinster House in the 1880s, contain the National Library and National Museum of Ireland. Merrion Square, immediately to the east, and Fitzwilliam Square, to the south, are two of the great 18th-century squares.

The oldest and largest of the city's squares is St. Stephen's Green, which was recorded in 1224 as common grazing land. It was enclosed and bordered with houses in 1663, although the imposing mansions now surrounding it were built principally in the 18th century. By 1887 the parkland was run down, and the Guinness family, whose former residence on the south side of St. Stephen's Green now houses the Department of Foreign Affairs, paid for its rehabilitation. St. Stephen's Green

From the western side of St. Stephen's Green to the river, and from there up the northern bank to Parnell Square,



The River Liffey, Dublin, with the dome of the Four Courts at right.
Bord Failte Photograph

runs the city's north–south axis. Grafton Street, long Dublin's premier shopping district, was designated for pedestrians only in the 1990s and has become a lively thoroughfare filled with street entertainers. It emerges onto College Green between the University of Dublin (Trinity College) and the 1729 Parliament House, which is now the Bank of Ireland's headquarters. Trinity College is Ireland's oldest university, founded in 1592, though many of its most distinguished buildings date from the 1700s.

Along the quays of the River Liffey are many monumental buildings, including James Gandon's Neoclassical masterpieces of the Custom House (1781–91) to the south and the Four Courts (1786–1802) to the north. The Custom House was burned in 1921 during the war of independence by republicans who wished to destroy British administrative records; the Four Courts was reduced by shellfire and mines at the outbreak of civil war in June 1922. Both have since been rebuilt by the government.

O'Connell Street—first called Drogheda and then Sackville Street—is Dublin's "downtown," an assemblage of shops, cinemas, and snack bars. The only building of any distinction to survive the warfare that swept the street in 1916 and again in 1922 was the General Post Office, headquarters of the 1916 rebellion. Badly damaged in the uprising, it was reconstructed behind its surviving 1815 classical facade in 1929. Opposite the Post Office stood Nelson's Pillar, a landmark for generations of Dubliners. It was built in 1808 by public subscription but was mysteriously blown up late one night in 1966. In 2000 Dublin Corporation began upgrading both the street and its shops.

At the top of O'Connell Street, Bartholomew Mosse constructed his Rotunda Hospital, also known as the Lying-In, which remains a maternity hospital. To support the hospital, he added a pleasure garden, assembly rooms, and a concert hall. Part of the assembly rooms now serves as the historic Gate Theatre. Behind the hospital is Parnell (formerly Rutland) Square, built in 1750. Many of its original Georgian houses are still intact. One, built for the Earl of Charlemont in 1762–65, now houses the Municipal Gallery of Modern Art. The Roman Catholic Pro-Cathedral was built on Marlborough Street, east of O'Connell Street, in 1816. Owing to their antagonism toward Catholicism, the municipal authorities refused to allow the cathedral to be erected on the main thoroughfare.

The 18th-century city commissioners circumscribed the new city with the North and South Circular roads. On Synge Street, close to the South Circular Roads, is the birthplace of the dramatist George Bernard Shaw. North of these peripheral streets the Grand Canal was constructed; to the south of them is the Royal Canal. Both entered the Liffey at the harbour entrance, and both connect with the River Shannon, though only the Grand is now navigable.

Phoenix Park

Dublin's Phoenix Park is the largest enclosed urban park in Europe. With a circumference of 7 miles (11 kilometres), it covers nearly 30 square miles (80 square kilometres) on the north bank of the Liffey. During the first visit by a reigning pontiff to Ireland, in September 1979, the religious service conducted by Pope John Paul II in the park attracted the largest gathering ever recorded in the country. Initially a royal deer park, Phoenix Park was opened to the public in 1747. Its zoo, celebrated for lion breeding, opened in 1831 and effectively doubled its size in 2001 when the African Plains section opened on land donated to the zoo by the president of Ireland. Nearby is Islandbridge, the site of World War I memorial gardens designed by Sir Edwin Luytens.

THE PEOPLE

Demography. During the second half of the 20th century, the population of Dublin and the surrounding area grew annually by only about 1 percent. Initially the trend in migration was from the countryside to the city. During the last quarter of the 20th century, however, central city areas began to lose population while new suburbs southwest and north of Dublin grew. Urban regeneration at the end of the 20th century attracted new dwellers to the inner city.

Religion. The administrative bodies of Ireland's main religious groups are based in Dublin. Although overwhelmingly Roman Catholic, Dublin is the most religiously diverse part of the country. The non-Catholic population has steadily declined since 1922, but Dublin still holds most of the Anglicans in the republic, as well as half of the Presbyterian clergy. Many older suburbs around the southern rim of the city form the "Protestant belt" of Dublin. An increase in the number of persons who profess other religious creeds reflects the growth of evangelical and charismatic Christian groups during the 1970s, and the number of Dubliners professing no religion, especially among the young, has increased.

THE ECONOMY

Industry. Dublin's major traditional industries—brewing, distilling, food processing, and textile manufacturing—all suffered a decline beginning in the 1970s, and this led to inner-city blight. The recession of the 1980s also led to a slump in the building trades. Several industrial estates, however, have been built in the suburbs around the city and, with the help of government grants, have attracted new enterprises, notably computers, electronics, chemicals, and engineering.

Economic resurgence

Finance and commerce. Dublin is the headquarters for Ireland's chief financial and commercial institutions. The economic pace has quickened markedly since 1973, when the country joined the European Economic Community

(EEC; in 1993 renamed the European Community in the European Union [EU]). In addition to the five major clearing banks, which have their main offices in Dublin, there has been a rapid increase in the number of other banks, principally from EU countries. The Irish Stock Exchange, an integral part of the British Stock Exchange system, is also located in central Dublin and is one of the oldest exchanges in the world, trading continuously since 1793.

Traffic through the port of Dublin has dwindled considerably, but an International Financial Services Centre in the former dock area near the Custom House, under the Custom House Development Authority, was set up in 1986. This venture reflected the commitment of Dublin's commercial and financial interests to plans for the single European market, with its attendant abolition of duties and tariffs within the EU.

Transportation. In 1986 Parliament reorganized transport for the capital by establishing Dublin Bus (Bus Átha Cliath) as a subsidiary of Córas Iompair Éireann (CIE), the national transport company. Another subsidiary of CIE, Irish Railways (Iarnród Éireann), provides suburban services and intercity connections with the rest of the country and Northern Ireland. Dublin's international airport is just north of the city at Collinstown.

ADMINISTRATION AND SOCIAL CONDITIONS

National and local government. Dublin is the administrative capital of the republic of Ireland, serving as the headquarters for government departments, their advisory committees, and associated agencies. The two houses of the Irish Parliament, the Dáil and the Seanad (Senate), meet at Leinster House in the centre of the city. The judiciary is based at the Four Courts. More than 40 countries have resident embassies, and several others are represented by honorary consuls. Just under one-third of the Irish electorate lives in the Dublin area's 11 constituencies, which are represented by 48 members of the proportionally elected Dáil.

Locally, three elected authorities administer Dublin: Dublin County Council, Dublin Corporation (for the city), and Dún Laoghaire Corporation (for the port of Dún Laoghaire, a separate borough to the south of the city). Although certain functions are reserved for the elected bodies, city and county managers perform the executive functions. Through the Local Appointments Commission, the state's Department of the Environment names the managers.

Health. Health care services are administered by the Eastern Regional Health Authority (formerly the Eastern Health Board), the largest of the republic's regional health boards. Health care is free, subject to a means test.

Trinity College

Education. The oldest of Dublin's universities is Trinity College (1592), the only college in the University of Dublin. For centuries Trinity was regarded as a bastion of the Protestant Ascendancy in Ireland, and for many years Roman Catholics were barred from taking degrees, though they could still attend. The ban was lifted in 1793 with the passage of the Catholic Relief Act, but Catholics were still not eligible for the college's full benefits until 1873. University College Dublin, a constituent college of the National University of Ireland, is the largest campus in Ireland, with more than 10,000 students. In 1989 the newest university, Dublin City University, was created from the National Institute for Higher Education at Dublin. Also in the capital are a number of other higher educational institutions, including colleges of technology, teacher-training colleges, and specialized vocational colleges.

CULTURAL LIFE

Dublin played a leading role in the cultural renaissance that began in 1884 with the establishment of the Gaelic Athletic Association for the revival of Gaelic games. It was broadened in 1893 with the foundation of the Gaelic League (Conradh na Gaeilge), which maintains its aim of promoting Irish language and folklore. The National Gallery, the Irish Museum of Modern Art, the Project and City Arts Centres, and many privately owned galleries reflect the liveliness of the visual arts. The area known as Temple Bar—bounded by Dame Street, Westmoreland Street, Parliament Street, and the Liffey—has been developed with a mix of boutiques, galleries, and studios.

The Abbey Theatre

Theatre and music. Early in the 20th century the cultural renaissance gained strong momentum in Dublin with the opening of the famous Abbey Theatre, an enterprise associated particularly with the poet William Butler Yeats and the playwrights John Millington Synge and Lady Gregory. In addition to producing their works, the Abbey later staged the first performances of Sean O'Casey's major plays. The old theatre burned down in the early 1950s, and with government aid a new theatre was opened in 1966 housing the main Abbey stage and the experimental Peacock Theatre. In 1928 Micheál MacLiammóir and Hilton Edwards started the Gate Theatre Company, which continues to flourish. The state-sponsored Arts Council, headquartered in Dublin, subsidizes the Abbey and Gate theatres and a number of small theatrical groups in Dublin.

The city's two main commercial theatres are the Gaiety, which stages annual opera seasons, and the Olympia. In 1980 the National Concert Hall was opened, and, after decades of unsuccessful attempts, the capital finally had a major concert venue. Radio Telefís Éireann, the national radio and television station, is also based in Dublin. It employs the country's principal symphony orchestra.

The city has produced a number of internationally famous pop musicians, including Sinead O'Connor and the postpunk band U2.

Publishing. The country's principal newspapers and periodicals are based in the capital. Dublin has two national daily papers, one evening paper, and three Sunday papers. A number of small but influential literary and current affairs magazines are published, both in Irish and in English. Since 1970 there has been an increase in the number of publishing houses devoted to literature, especially poetry.

Sports. Phoenix Park no longer holds races, but racing flourishes at Loughlinstown and Fairy House. There is also a greyhound track at Harold's Cross. The traditional Gaelic games—hurling and Gaelic football—are played at Croke Park, on the north bank of the Royal Canal. International rugby and association football (soccer) matches are held at Lansdowne Road, and Belfield at University College Dublin attracts major competitions. Golf is a popular sport.

History

FOUNDATION AND EARLY GROWTH

From prehistoric times people have dwelt in the area about Dublin Bay, and four of Ireland's five great roads converged near the spot called Baile Átha Cliath, the name stamped by Dublin's postmark. Dublin appeared in Ptolemy's *Geōgraphikē hyphēgēsis* ("Guide to Geography"; *c.* AD 140), and 151 years later "the people of Dublin," it was recorded, defeated an army from the province of Leinster. Yet, despite indications of habitation there 2,000 years ago, the first settlement for which one can discover any historical proof was not Celtic, but Norse.

The Vikings, or Norsemen, invaded in the 9th century (*c.* 831) and built upon the river's south bank on the ridge above, where Dublin Castle rose 400 years later. The Vikings beat off most Irish attacks until 1014, when they were defeated at the Battle of Clontarf on the north shore of the bay. They nevertheless reoccupied the town, and Viking Dublin survived and grew, though eventually the Norse kings were reduced to earls under Irish overlords.

Viking Dublin

In 1167 the Norsemen supported Roderic (Rory O'Connor) of Connaught (Connacht), claimant to the high kingship of Ireland, in driving into exile their overlord Dermot MacMurrough, king of Leinster. Dermot returned in 1170 with an army of Anglo-Normans from Wales and retook Dublin. Alarmed lest his Anglo-Norman vassals should claim Ireland for their own, King Henry II of England hurried over with an army to affirm his sovereignty. This action proved to be the key to Dublin's development, for it was to establish the site as the centre of government.

Until the middle of the 17th century, Dublin remained a small, walled medieval town, dominating only the Pale—the thin strip of English settlement along Ireland's eastern seaboard. In the 500 years to 1660, three uprisings in the city were suppressed, and a Scottish invasion and the ravages of the Black Death were endured.

At the time of the Reformation, Dublin had become Protestant. During the English Civil War the city's royalist defenders, after contemplating joining forces with an armed Irish Catholic confederacy, surrendered the city in 1649 to Oliver Cromwell's English parliamentary army. By the end of the Cromwell era, Dublin was a town of only 9,000 inhabitants. The turreted city wall with its eight gates was a shambles; the two cathedrals tottered; and the dilapidated castle was, as Cromwell himself put it, "the worst in Christendom." Yet, in the 18th century, Dublin was to become the second city of the British Empire.

ASCENDANCY IN THE 18TH CENTURY

The city's remarkable resurgence began at the end of the 17th century, when thousands of refugee Huguenot weavers from France settled in Protestant Dublin after the revocation of the Edict of Nantes, in 1685, curtailed their privileges. Flemish weavers came in their wake, and soon the cloth trades were flourishing. It was not long before Dublin's competition with English cloth interests prompted the British Parliament to impose export restrictions.

In the course of the 18th century, economic prosperity led to the development of Georgian Dublin. Development spread beyond the old medieval walls; more bridges were erected over the Liffey; and splendid new suburbs arose to the north and east. The city that emerged was, in essence, that of the Dublin of today.

The Georgian period

Culturally, the century was one of the richest periods in the city's history. Jonathan Swift was dean of St. Patrick's Cathedral between 1713 and 1745, and other noted literary figures—Oliver Goldsmith, Sir Richard Steele, and William Congreve—were active in Dublin. In the New Musick Hall, Handel conducted the first public performance of his *Messiah* in 1742. For members of the Protestant Ascendancy, as the English establishment was called, Dublin was a gay, fashionable city of elegance and wit.

It was something less than that, however, for Roman Catholics, who constituted the majority of the population. At the beginning of the century the Irish Parliament, dominated by the Protestant Ascendancy, passed the Penal Laws, a series of harsh discriminatory measures against the Catholics of Ireland. These laws disfranchised Catholics, placed restrictions on their ownership of property, hindered them from entering the professions, and obstructed Catholic education. The majority of the population was kept in extreme poverty and degradation.

EVOLUTION OF THE MODERN CITY

In 1801 the Act of Union between England and Ireland abolished the Irish Parliament and drastically reduced Dublin's status. With no governmental duties to compel their presence in Dublin, the leading figures of the Ascendancy returned to England. The city fell into a decline from which it recovered only 150 years later. Dispossessed farmers crowded into the tenantless Georgian houses, reducing these once elegant structures to slums. Anyone who owed more than 10 shillings could be imprisoned, and, until the legislation was revised in 1864, Dublin's jails overflowed with debtors.

With the easing of the Penal Laws, however, a Roman Catholic middle class emerged, sending its sons to university and into the professions. In 1829 the political dexterity of the Irish Catholic lawyer Daniel O'Connell achieved passage of the Emancipation Act, repealing the Penal Laws and enabling Catholics to sit once again in the British Parliament. After reforms in Dublin's municipal government, O'Connell became, in 1841, the first Roman Catholic mayor of the city since the 17th century. For the first time in 200 years Roman Catholic churches and schools were built, and in 1854 the Catholic University of Ireland (now University College Dublin) opened on St. Stephen's Green, with John Henry Newman as rector.

The railways came to Ireland in 1834, when a seven-mile link connected Dublin with the port of Kingstown (now Dún Laoghaire). As a result, suburbs began to grow up along the coast to the south. Suburban development around the city continued and intensified over the next 70 years.

Although Dublin remained modestly prosperous on the surface, it was festering underneath. The city had some of the worst slums in Europe. Infant and child mortality rates were uncommonly high, with tuberculosis constituting a particular scourge; sanitation and hygiene were practically nonexistent. An investigation in 1910 revealed that 20,000 families were each living in only one room. A two-week survey of 22 public houses, or taverns, disclosed more than 46,000 women and 28,000 children among the customers.

As the 20th century opened, political tensions increased. In 1914 the Home Rule Party secured home rule for Ireland from the government of the United Kingdom; but within months, when World War I erupted, the agreement was suspended. For some years before the outbreak of the war, the Irish Republican Brotherhood (Fenians), who had been quiescent since the failure of their rebellion in 1867, had been secretly reorganizing. When war came, they made plans for another rebellion against the British. With the help of the Irish Citizen Army, a small volunteer workingmen's corps, and the Irish Volunteer Army, a rising took place on Easter Monday 1916. Leaders of the movement proclaimed an Irish republic. The rebels occupied public buildings in the centre of the city, which they held for a week. Commerce and industry came to a halt, and a quarter of the city's population of 390,000 went on public relief.

The Troubles of 1916–23

Finally defeated, the rebels were marched through the streets of Dublin to the jeers and abuse of the populace. But the establishment of martial law in Dublin, the execution of the leaders within 10 days, and the mass imprisonment of those thought to be implicated in the uprising roused Irish public opinion as the rebellion itself had not. Guerrilla warfare by the Irish Republican Army spread through the country in 1919, continuing through two years of terror and counterterror. Dublin was one of the worst affected areas in Ireland and for much of those two years was subject to martial law.

A treaty was concluded in 1921 establishing the Irish Free State, but an antitreaty contingent of the republican army protested and took possession of the Four Courts building. The rebels were eventually driven out by artillery, an event that marked the start of 11 months of murderous civil war between the factions that were for and against the treaty. Once again Dublin suffered heavily in the conflict. The end of the civil war in 1923 did not mean the end of gunfire in the streets, however. Political assassinations and armed raids continued until the early 1930s, and hostilities remained a marked feature of Dublin life.

AFTER NATIONAL INDEPENDENCE

Between 1922 and 1932 the first administrations of the new Irish Free State were preoccupied with trying to establish new government institutions and to repair the damage inflicted on the economy by the Troubles of 1916–23. Housing took a low priority, and it was not until the advent of Eamon De Valera's Fianna Fáil government in 1932 that a concerted program of home building got under way. Some of the worst inner-city slums were cleared, and the people moved to new housing projects on the outskirts of the city. With the introduction of better health care, old-age pensions, and children's allowances, the position of Dublin's poor began to improve.

With the outbreak of World War II, housing construction came to a halt because of a shortage of building material, much of which was imported. Since Ireland remained neutral, Dublin escaped the worst effects of the war, although there were isolated bombing incidents. Food, with some exceptions, was plentiful, but the scarcity of gasoline made private transport nonexistent and severely limited public transport. Politically, Dublin had the mysterious atmosphere of other neutral "whispering galleries" like Madrid and Lisbon, heightened by the presence of both Allied and Axis diplomats.

Dublin during World War II

After the war, as shortages eased, the city began again to spread into the surrounding countryside, and more suburbs took shape. In 1969, high-rise apartment blocks were built in the new satellite town of Ballymun; unfortunately, Ballymun proved no more immune than other places in Europe and the United States to social problems like crime and vandalism that attend such buildings. The situation

there aroused criticism that the design of the tower blocks was unsuitable for family living.

This surge in building was a symbol of the prosperity that rejuvenated the city in the 1960s and '70s. Tourism became a major industry, and Ireland's membership in the EU brought more political, economic, and cultural organizations to Dublin. Development slowed with the onset of the economic recession in the early 1980s but quickened again as the economy improved later in the decade. By the mid-1990s what became known as the "Celtic Tiger" was flourishing economically, which led to a further revival in Dublin.

The social and economic changes that have come about since 1945 inevitably put pressure on historic Dublin, and there is an energetic conservation movement. In 1988 the city celebrated its millennium, arousing much thought and comment about Dublin's past and future, especially concerning the quality of its urban life. The city's regeneration was recognized in 1993, when it was designated as the European City of Culture.

BIBLIOGRAPHY

Physical and human geography: JOHN HARVEY, *Dublin, a Study in Environment* (1949, reprinted 1971), offers a general description. V.S. PRITCHETT, *Dublin: A Portrait* (1967), is an introduction, capturing the character of the city. On city planning, see MICHAEL J. BANNON (ed.), *The Emergence of Irish Planning, 1880–1920* (1985). PETER WYSE JACKSON and MICHELINE SHEEHY SKEFFINGTON, *Flora of Inner Dublin* (1984), is an illustrated study. Historic sites and buildings are presented in ADRIAN MACLOUGHLIN, *Guide to Historic Dublin* (1979). The literary landmarks of the city are discussed in VIVIAN IGOE, *Literary Guide to Dublin* (1995); PATRICIA HUTCHINS, *James Joyce's Dublin* (1950); and JACK MCCARTHY, *Joyce's Dublin: A Walking Guide to Ulysses* (1986). Other guide books include CAROL BARDON and JONATHAN BARDON, *If Ever You Go to Dublin Town: A Historic Guide to the City's Street Names* (1988). ALEXANDER J. HUMPHREYS, *New Dubliners: Urbanization and the Irish Family* (1966), is a sociological analysis. A case study of Dublin is included in CHRISTOPHER T. WHELAN and BRENDAN J. WHELAN, *Social Mobility in the Republic of Ireland: A Comparative Perspective* (1984). Current social and economic developments are discussed in the *Administration Yearbook and Diary,* an annual publication of the Institute of Public Administration. Social life and customs are examined in KEVIN CORRIGAN KEARNS, *Dublin's Vanishing Craftsmen: In Search of the Old Masters* (1987); and JOHN O'DONOVAN, *Life by the Liffey: A Kaleidoscope of Dubliners* (1986). RICHARD ELLMANN, *Four Dubliners: Wilde, Yeats, Joyce, and Beckett* (1987), explores literary traditions. JAMES KILLEN and ANDREW MACLARAN (eds.), *Dublin: Contemporary Trends and Issues for the Twenty-first Century* (1999), debate the city's future.

History: JOHN THOMAS GILBERT, *A History of the City of Dublin,* 3 vol. (1854–59, reprinted 1978), is comprehensive. PETER SOMERVILLE-LARGE, *Dublin: The First Thousand Years* (1988), is a modern historical survey. Other histories include GEORGE A. LITTLE, *Dublin Before the Vikings: An Adventure in Discovery* (1957); CHARLES HALIDAY, *The Scandinavian Kingdom of Dublin,* 2nd ed. (1884, reprinted 1969); JOHN PENTLAND MAHAFFY, *An Epoch in Irish History: Trinity College, Dublin, Its Foundation and Early Fortunes, 1591–1660* (1903, reprinted 1970); R.B. MCDOWELL and D.A. WEBB, *Trinity College, Dublin, 1592–1952: An Academic History* (1982); MAURICE CRAIG, *Dublin, 1660–1860* (1952, reissued 1980); CONSTANTIA MAXWELL, *Dublin Under the Georges, 1714–1830,* rev. ed. (1956); and MARY E. DALY, *Dublin, the Deposed Capital: A Social and Economic History, 1860–1914* (1984). JIMMY WREN, *The Villages of Dublin,* enlarged ed. (1987), provides a survey of the history of the newer suburbs. The intellectual, cultural, and political history of the 19th century is surveyed in RICHARD M. KAIN, *Dublin in the Age of William Butler Yeats and James Joyce* (1962, reissued 1972); and TERENCE BROWN, *Ireland: A Social and Cultural History, 1922 to the Present* (1985). PETER SHERIDAN, *44: A Dublin Memoir* (1999), recalls the city over the second half of the 20th century. (B.E./J.O'B.R./Ed.)

Dutch Literature

Of the earliest inhabitants of the Netherlands, only the Frisians have survived, and they have maintained a separate language and literature since the 8th century. The remainder of the Netherlands was colonized by the Saxons and Franks between the 3rd and 9th centuries, resulting in a predominantly Frankish culture in the south and Saxon or an amalgam of Saxon and Frankish language and culture elsewhere.

Under the less nomadic Franks, the south prospered more than the north, and there a literary language first developed. Because of marked differences between the dialects of the east, the centre, and the west (Flanders, with features that linked the coastal dialects with Old English), the development was very gradual. In the early Middle Ages, when Latin and, later, French were the languages of the educated, the vernacular was largely confined to unrecorded oral legend and folk songs. The earliest text that can claim to contain examples of Old Dutch was the early 10th-century "Wachtendonck Psalm Fragments."

This article is divided into the following sections:

Medieval literary works 566
  Poetry and prose
  Songs, drama, and the rhetoricians
The Renaissance and Reformation 566
The 17th century 567
  The writers of the "Golden Age"
  Religious poetry
The 18th century 567
The 19th century 567
  Romanticism
  Movement of the 1880s
The 20th century 568
Bibliography 568

MEDIEVAL LITERARY WORKS

Poetry and prose. The work of Heinrich von Veldeke, the earliest known poet to use a Dutch dialect, typified the age's religious zeal, which emanated from the French centres of learning. In addition to his *Eneit* (*c.* 1185), a chivalrous rendering of Virgil's *Aeneid,* and his love lyrics, which were important for German poets, Heinrich produced *Servatius,* a saint's life written in the Limburg dialect. Dutch 13th- and 14th-century texts were generally written in the cultural centres of Flanders and Brabant, where, for reasons of trade, the prevailing influence was French. Throughout Europe the Crusades brought courtly romances into vogue, and Dutch romances, following French models, were written about events from classical history, such as Segher Diergotgaf's *Paerlement van Troyen* ("Parliament of Troy"); about Oriental subjects; or, most popular of all, on themes from Celtic sagas, including the Arthurian cycle. But by the 1260s chivalry was on the decline; the titles of Jacob van Maerlant's later works bear witness to a late 13th-century reaction against romance. Van Maerlant's compendia of knowledge, including his *Der naturen bloeme* ("The Flower of Nature") and *Spieghel historiael* ("The Mirror of History"), answered a demand for the kind of self-instructional literature that long remained a characteristic of Dutch literature. The change in social patterns at this time is also evident in two epic tales. *Karel ende Elegast* ("Charles and Elegast"), probably an original Flemish chanson de geste of the 12th or 13th century, describes with feudal reverence Charlemagne's adventures in the magic world of folklore. *Van den vos Reinaerde* (*c.* 1240; "Reynard the Fox") is the Flemish poet Willem's version of a translation by another Fleming, Aernout, of the French *Le Plaid,* which, by contrast, brilliantly satirizes feudal society and the epic manner.

Mystical writing reached a remarkable lyrical intensity in the poetry and hortatory prose of a Brabantine laywoman, Hadewijch (late 12th, early 13th century), and this inspired later mystics, greatest of whom was Jan van Ruysbroeck, a disciple of the German mystic Meister Eckehart and the Netherlands' greatest medieval prose writer. His most important work was *Die chierheit der gheestliker brulocht* (1350; *The Adornment of the Spiritual Marriage,* or *The Spiritual Espousals*), concerning the soul in search of God. His work was part of a renewed ecclesiastic concern to instruct the laity, which resulted in a wealth of Bible stories, legends, and didactic short stories. Of these, *Beatrijs,* an early 14th-century Flemish verse rendering of a popular legend, is told with such humanity and restraint that it still inspires modern versions (*e.g.,* those by Maurice Maeterlinck and Pieter Cornelis Boutens). Mystical writing

Songs, drama, and the rhetoricians. The earliest recorded songs suggest a Germanic rather than a Romance tradition. Because the first extant plays—the 14th-century *Abele spelen* ("seemly plays")—were entirely secular (and may have been the first of such in Europe), incorporating romantic themes from the earlier songs, there is reason to attribute the emergence of drama in the Netherlands as much to mime and song as to liturgical action. The only evidence of early liturgical drama is the Latin *Officium stellae* of the 14th century, after which there is nothing until 1448–55, when the play cycle on the seven joys of Mary was first performed at Brussels. Of the many miracle and morality plays, two deserve special mention: *Mariken van Nieumeghen* (late 15th century; "Mary of Nijmegen") and *Elckerlyc* (of about the same date). The first anticipates the Renaissance in its psychology and treatment; the second, entirely medieval in its conception, is the original of the English *Everyman.* Both were written by members of *rederijkerskamers,* or chambers of rhetoric, institutions that spread from the French border in the 15th century. Organized like guilds, with functions similar to those of the French medieval dramatic societies, the chambers were commissioned by the town protecting them to provide the ceremonial and entertainment at religious and secular festivals, and they were influential in popularizing art and morals. Drama by this time was in the hands of the laity rather than the church, and the introduction of secular themes made it necessary to perform outside of religious buildings, using stages or carts. The survival of the chambers depended on literary performance, and members organized national festivals and competitions. A record of one such festival, held in 1561, is the illustrated *Antwerps landjuweel* (1562; "Antwerp National Contest"). The chambers of rhetoric

THE RENAISSANCE AND REFORMATION

The literature of Flanders and Holland must be considered as a whole until about 1585, when the fall of Antwerp marked the final rift between the Protestant north and the Roman Catholic south. The new art of the Renaissance, coming to the Netherlands from Italy through France, first found expression in writers such as Lucas de Heere, who had fled from the Catholic southern provinces for religious reasons. Chapbooks, containing prose versions of medieval romances, folk songs, and *rederijkers* ("rhetoricians") verse; Reformation propaganda; marching songs of the Calvinist revolt against Spain; these and the first sonnets, the first dissertations in the vernacular, and the first grammars of the Dutch language displayed the restlessness of an age of change. So, while the Catholic Anna Bijns was fulminating against Lutheranism in her glowing satirical verse, which was countered later by the Calvinist Marnix van Sint Aldegonde in his polemical attack on the Catholic Church, the echoes of classical antiquity were reaching the Netherlands in the odes, sonnets, and translations of Jan Baptista van der Noot and Jan van Hout. Carel van

Mander, painter and poet, introduced scholarly vernacular prose writing, though the Latin prose of Erasmus had been famous throughout Europe for nearly a century.

Van der Noot's Petrarchan sonnets, written in the manner of the French poet Pierre de Ronsard, were published in London, where he was then in exile for participating in an insurrection in 1567. The two great moderates of the age were the Erasmians Henric Laurenszoon Spieghel and Dirk Volkertszoon Coornhert, liberal Humanists who espoused a social, undogmatic Christian ethic. Spieghel's poetry is generally more intellectual than Coornhert's prose, which was influenced by Montaigne and the Bible, with a remarkably supple and lucid, even entertaining, style. It was Coornhert and his successors, in particular the translators of the Dutch authorized version of the Bible (published in 1637), who laid the foundations of the standard language.

THE 17TH CENTURY

While the Spanish hold on the Catholic south of the Netherlands during and after the Eighty Years' War (1568–1648) caused a decline in Brabant and Flanders, there was a spectacular expansion in Holland, to which artists, intellectuals, and financiers had fled from the Spanish armies. The emergence of Amsterdam and The Hague as capitals of an empire and the birth of civic pride in writers of the "Golden Age" symbolized the final passing of a medieval age belonging to Ghent, Bruges, Liège, and Antwerp.

The writers of the "Golden Age." Spieghel, the greatest of a generation straddling the old and the new, wrote for both the burgher and scholar. His *Nieujaarliedekens* ("New Year Songs") and *Lieden op 't Vader Ons* ("Songs on the Lord's Prayer") continued a medieval tradition in a Renaissance style echoing Erasmian moderation; his learned *Twe-spraack vande Nederduitsche letterkunst* (1584; "Dialogue on Dutch Literature") was intended to popularize the use of a national language. His most scholarly work, the unfinished *Hertspieghel* (1614; "Mirror of the Heart"), was particularly abstruse because it represented a first attempt at philosophizing in the vernacular and in poetry.

The dichotomy inherent in the Renaissance—between popular religious revival and Humanism—was particularly marked in Holland because of the incompatibility of Calvinistic principles with the ideals of pagan antiquity. This caused a tense ambivalence in many writers of the 17th century who took both their religion and their art seriously. Daniël Heinsius, a celebrated Humanist at the University of Leiden, wrote plays in Latin, but he also contributed to the vernacular by writing *Hymnus ofi lof-sanck van Bacchus* (1614; "Hymn in Praise of Bacchus") and an equally devout *Lof-sanck van Jesus Christus* (1615).

A poet, playwright, and painter, Gerbrand Adriaenszoon Bredero took his material from the life of the commoner; his medium was the folk song, farce, or comedy. His secular songs in medieval style and devotional songs in Renaissance verse told of a passionate devotion to women and yearning for religious moderation. While his three tragicomedies were not successful, his three farces marked the zenith of the medieval genre. Contemporary life in Amsterdam provided material for his two comedies, including his masterpiece, *Spaanschen Brabander* (performed 1617).

Amsterdam was the home of the poet and dramatist Joost van den Vondel. Like Bredero, he was self-educated, and he resolved the conflict between artistic and religious leanings only when he entered the Roman Catholic Church at age 54. This was a courageous act of faith at a time when Catholics formed an unpopular minority. It is a measure of van den Vondel's indomitable personality that his attitude toward contemporary people and events, of which he was a fearless chronicler, still prevails even when history has recorded a different view. His plays, however, are too austere for modern readers, although, in his Sophoclean *Jeptha* (1659) and his Baroque masterpieces *Lucifer* (1654) and *Adam in ballingschap* (1664; *Adam in Exile*), he was as great an artist of the Counter-Reformation as his contemporary the Flemish painter Peter Paul Rubens.

Italian Baroque influence

The aristocratic Pieter Corneliszoon Hooft was one of a fortunate few in Holland to bring the refinements of the new art directly from Italy. He lavished an Italianate flourish on his sonnets and plays, the studied prose of his letters, and a monumental unfinished history of the war against Spain. His castle at Muyden became a thriving centre for the entertainment of artists and scholars attracted not only by a mutual interest in poetry, music, and learning but also by the charm of such gifted young women as the Roemer Visscher daughters, Anna and Maria.

Anna Visscher in verse, like her father Roemer in prose, popularized ethics in a manner that was to bring Jacob Cats unmerited fame. Cats's prolix moralizing, pedestrian doggerel, and patronizing tone forced their way into his country's literature if only because of the disastrous influence they had on the taste of their middle-class readership.

A more harmonious individual, Constantijn Huygens, had all the qualities to which Dutchmen of his day might aspire. A man of strict Calvinist principles, he was an able diplomat who wrote trenchant, shrewd, and witty verse and made excellent translations of John Donne's poetry.

Religious poetry. Three clerics contributed religious verse of considerable merit. The Roman Catholic Joannes Stalpaert van der Wiele wrote *Den schat der geestelycke lofsangen* (1634; "The Treasury of Devotional Praise"), containing songs of medieval simplicity and devotion. Jacobus Revius, an orthodox Calvinist, was a master of the Renaissance forms and the sonnet. Ironically, Dirk Rafaëlszoon Camphuysen, removed from his parish because of his unorthodoxy, satisfied a widespread demand for personal, devotional poetry in *Stichtelycke rymen* (1624; "Edifying Poems"). Equally popular was the introspective mystical poetry by the ascetic Jan Luyken, a layman who began by writing hedonistic songs in *De Duytse lier* (1671; "The Dutch Lyre"), containing fine love lyrics.

THE 18TH CENTURY

The appearance in 1669 of the first literary society (*dichtgenootschap*) was an omen of a decline in Dutch literature lasting through the 18th century. Material well-being sapped the vitality of the nation. Even the talented poet Hubert Poot suffered from the delusion of his day that rococo flourish and prescribed form were the criteria of poetry. Prose, too, consisted almost exclusively of translations and bombastic disquisitions. Significantly, Justus van Effen wrote in French before he founded *De Hollandsche spectator* (1731–35). The simple style of his moralizing essays contrasts with the work of his contemporaries, and his descriptive realism links him with two popular Dutch authors, Betje Wolff (byname of Elizabeth Wolff-Bekker) and Aagje Deken (byname of Agatha Deken).

Betje Wolff, essayist and poet, blended rationalism and romanticism in her creative genius. Her association with Aagje Deken as friend and fellow writer produced the classic epistolary novel *De historie van mejuffrouw Sara Burgerhart* (1782; "The History of Miss Sara Burgerhart"), dedicated to "Dutch young ladies." Remarkable for its wit and realism, it owed much to the English novelist Samuel Richardson. Wolff's intelligence and humour also dominated the original didactic purpose of the pair's eight-volume *Willem Leevend* (1784–85).

The first Dutch novel

By the end of the century a number of poets—including Hieronymus van Alphen, Rhijnvis Feith, Jacobus Bellamy, and Antony Staring—were reacting against Neoclassicism. The most admired and influential poet of the period was Willem Bilderdijk, whose versatile genius was almost smothered with excesses of rhetoric but whose Protestant zeal had repercussions in the Réveil (Revival), a Calvinist fundamentalist movement that gave impetus to the literary revival of the 1830s.

THE 19TH CENTURY

Romanticism. Although Jacob Geel's essays in *Onderzoek en phantasie* (1838; "Inquiry and Fantasy") set a new standard in philological and philosophical criticism in Dutch literature, Geel's liberal rationalism was almost swept aside by the growing wave of Romanticism. Simultaneously, the freethinking born of the Enlightenment roused the militancy of the Calvinists, who realized that their entrenched position was being threatened. Willem Bilderdijk and his disciple Isaäc da Costa reminded the nation of its divine mission, and foreign historical novels

(particularly the work of Chateaubriand and Sir Walter Scott) provided models for historical national Romanticism. In 1826 David van Lennep published a paper calling for novels modeled on Scott; and his son Jacob was the first of many writers to respond, with *De pleegzoon* (1833; *The Adopted Son*). Aernout Drost, author of *Hermingard van de Eikenterpen* (1832; "Hermingard of the Oak Burial Mounds"), set at the beginning of the Christian Era, also started a new literary journal, *De muzen* (1834), which, like his novel, was true to the spirit of the Réveil. Two men on the journal's staff, a historian, R.C. Bakhuizen van den Brink, and a future leader of the literary revival, Everhardus Johannes Potgieter, continued the campaign to improve critical standards in *De gids* ("The Guide"), known as the "Blue Butcher" because of its merciless treatment of complacency. Potgieter defined the historical novel, and Anna Bosboom-Toussaint put his ideas into effect, transposing the universal Christian idealism of Drost to the national Protestant faith of the Golden Age. Bosboom-Toussaint's best known book, *Majoor Frans* (1874; "Major Francis"), was not historical, belonging rather to an era of liberal politics and female emancipationists.

Nicolaas Beets, although feted as a national Protestant poet, owes his enduring fame to his sketches in *Camera obscura* (1839), with their stylistic virtuosity and Dickensian observation of detail. Potgieter's allegorical humour was less direct in its appeal, and his quest for originality tended to deprive his style of simplicity and clarity. The perceptive and often scathing critic Conrad Busken Huet, a progressive who left the church, placed Dutch writing in a truer perspective with western European writing. His essays were collected in *Litterarische fantasiën en kritieken* (1868–88; "Literary Fantasies and Criticisms"), and his later work was best represented by *Het land van Rembrand* (1882–84). Meanwhile, a furor had been caused by an entirely unknown writer, Multatuli (pseudonym of Eduard Douwes Dekker), whose *Max Havelaar* (1860; Eng. trans. 1927), a satire of Dutch exploitation of the Dutch East Indies, unexpectedly revealed a stylistic innovator of genius. Dekker's writing, in *Wouterje Pieterse* (1865–77; Eng. trans. 1904) and *Minnebrieven* (1861; "Love Letters"), vibrated between extremes of sentimentality and anarchy, iconoclasm and utopianism. Although poetry as a convention was anathema to him, Dekker was greatly admired by the young men of the new generation, such as Jacques Perk, who wrote sketches in Dekker's humorous style before composing a sonnet cycle, *Mathilde* (published posthumously in 1882), which opened a new epoch in Dutch literature.

De nieuwe gids

Movement of the 1880s. The appearance of the periodical *De nieuwe gids* ("The New Guide") in 1885 marked the beginning of an important renaissance of literature in the northern Netherlands. Unlike the earlier periodical *De gids*, it pursued an exclusively aesthetic ideal. Leaders of the movement were the poets Willem Kloos and Albert Verwey and the violent and lyrical critic Lodewijk van Deyssel. Among others prominent in the movement were the dramatist, poet, and prose writer Frederik Willem van Eeden; Herman Gorter, who became the foremost poet after his poem "Mei" ("May") appeared in 1889; and the poets Pieter Cornelis Boutens and Jan Hendrik Leopold.

THE 20TH CENTURY

The writers of the Dutch revival of the 1880s were essentially individualistic, but in the next generation a new concern for philosophical and social problems became apparent. The poetry of a prominent socialist writer named Henriëtte Roland Holst-van der Schalk was characterized by a desire for justice and charity. The socialist dramas of Herman Heijermans were internationally successful. A group of Naturalist-Realist novelists—including Frans Coenen and, most gifted of all, Marcellus Emants—flourished. Arthur van Schendel made his debut with Neoromantic fiction, and Louis Marie Anne Couperus was at his peak as a stylish chronicler of life in The Hague.

Significant early poets were A. Roland Holst, J.C. Bloem, and P.N. van Eyck, a philosophical poet and essayist. Immediately after World War I two poets emerged: Hendrik Marsman, an advocate of free verse and representative of the Vitalist movement; and the pessimistic Jan Jacob Slauerhoff, whose works reflect the restless romanticism and disillusionment that characterized his life.

The literary periodical *Forum* was founded in 1932 by Menno ter Braak and Edgar du Perron, leaders of a movement that aimed to replace superficial elegance with greater sincerity and warned against the German threat before the war. The most important mid-20th-century Dutch writer, Simon Vestdijk, was originally associated with the *Forum* group, while Ferdinand Bordewijk's terse style produced hauntingly original fiction. The most original poet was Gerrit Achterberg, whose poems explore the boundary between life and death.

During the Nazi occupation, free literature either stopped or was published secretly. The poets known as *Vijftigers* (Men of the Fifties) rejected the reflective lyricism of the interwar years in favour of an experimentalist style that drew on Dada, Surrealism, and primitive and children's art to create a maximum of physicality. Minimal punctuation, neologisms, and startling associative imagery were used in the poetry of Lucebert (pseudonym of Lubertus Jacobus Swaanswijk). The poets of the following decade, among them J. Bernlef (pseudonym of Hendrik Jan Marsman), reacted with a deliberately low-key, "prosy" style. In Hans Cornelis ten Berge's richly allusive poems, the poet's ego is submerged in a closely structured exploration of language, myth, and history. In the 1970s Gerrit Komrij and others returned to traditional forms such as the sonnet and to rhyme, often with knowing and ironic undertones.

Postwar novelists

Postwar novelists showed the influence of the Nazi occupation in various ways. Anna Blaman treated existential solitude, while Willem Frederik Hermans' classically constructed stories and novels, notably *De donkere kamer van Damocles* (1958; *The Dark Room of Damocles*) and *Nooit meer slapen* (1966; "No More Sleep"), compellingly present a hostile universe that is chaotic and unfathomable. War, for Hermans, is simply an intensification of the abject human condition. Gerard (Kornelis van het) Reve, who made his debut as a deadpan chronicler of postwar malaise in *De avonden* (1947; "The Evenings"), concocted, in such books as *Nader tot U* (1966; "Nearer to Thee"), an extravagant and virtuoso blend of fact and fiction in the name of Romantic Decadence. Inventive panache, vitality, and philosophical reflection mark the fiction of Harry Mulisch, from *Het stenen bruidsbed* (1959; *The Stone Bridal Bed*), set in postwar Dresden, E.Ger., to a later treatment of the aftermath of occupation, *De aanslag* (1982; "The Attack"). Though he belongs chronologically to the war generation, Jan Wolkers began writing in the 1960s and brought a visual artist's sensibility to his often brutal stories and novels. Reactions to the painful loss of empire in the East Indies ran the gamut of nostalgia, affection, bitterness, and alienation in the work of Beb Vuyk (byname of Elizabeth Vuyk), Maria Dermoût, and Albert Alberts, and the colonial experience continues to be a source of inspiration. The tradition of sombre and anecdotal realism, dating from the late 19th century, was continued with great popular success by Maarten 't Hart, who in *De aansprekers* (1979; *Bearers of Bad Tidings*) drew fruitfully on the experience of a strict Calvinist upbringing. The "academic" school of writers associated with the magazine *De Revisor*, founded in 1973, preferred elegantly crafted analytical fictions to "mere" storytelling. Fictional miniaturism continued to thrive in short stories by Anton Koolhaas, Simon Carmiggelt, and F.B. Hotz.

It should be emphasized that from the 1930s Dutch literature and Flemish literature have been part of a composite literary culture: the writers, literary organizations, and departments of culture of the two countries have worked in close cooperation. For this reason the interested reader should consult the article BELGIAN LITERATURE: *Flemish*.

BIBLIOGRAPHY. C.G.N. DE VOOYS and G. STUIVELING, *Schets van de Nederlandse letterkunde*, 30th ed. (1966); GERARD KNUVELDER, *Handboek tot de geschiedenis der Nederlandse letterkunde*, 5th ed. (1970–76); THEODOR WEEVERS, *Poetry of the Netherlands in Its European Context, 1170–1930* (1960); JAMES ANDERSON RUSSELL, *Dutch Romantic Poetry* (1961); and REINDER P. MEIJER, *Literature of the Low Countries*, new ed. (1978).

(G.W.H./P.K.K./P.F.V./Ed.)

The Earth:
Its Properties, Composition, and Structure

If the Earth were reduced to a tabletop globe 50 centimetres (20 inches) in diameter, the portion accessible to direct observation through even the deepest mines and boreholes would be the equivalent of a very thin skin less than 1 millimetre (0.04 inch) thick. It is therefore not surprising that scientific investigators did not develop a picture of the Earth's interior until well into the 20th century, and only since the 1960s have they come to understand the dynamic processes that shape the terrestrial surface.

The Earth is a nearly spherical body with an equatorial radius of slightly more than 6,378 kilometres (3,963 miles). Compared with the other planets of the solar system, it is only of intermediate size and is substantially smaller than the giant planets Jupiter, Saturn, Uranus, and Neptune. The Earth's Moon, which has a radius of 1,738 kilometres, is relatively large in comparison to the Earth itself. This fact has been of great importance in determining the history of the Earth's rotation, for the Earth and the Moon raise tides in the bodies of one another, resulting in the dissipation of energy into heat, which in turn leads to the slowing of the Earth's spin velocity on its axis and the recession of the Moon. In fact, if the present rates of slowing and recession are linearly extrapolated backward in time, the Moon is found to have been impossibly close to the Earth at a time within the geologic record—a seemingly unexplainable paradox. Only very recently has it been shown that the present distribution of continents and oceans produces an anomalously high rate of slowing, and so the Moon need never have been extremely close to the Earth.

The outstanding feature of the Earth as a planet is the presence of liquid water. Water is vital not only for the biosphere but also for the geologic processes of erosion, transport, and deposition that shape the Earth's surface. Yet, if the Earth were closer to the Sun, the water would be vaporized; if farther, it would turn to ice. Two-thirds of the terrestrial surface is covered by oceans. It was long thought that the continents, constituting the remaining one-third of the surface, had been fixed in position throughout the Earth's history. Gradually some Earth scientists dared to suggest that there had been major continental displacements, and finally, during the 1960s, investigators developed the full picture of seafloor spreading and plate tectonics. The continents, though constantly in motion, are in fact the oldest portions of the Earth's surface, for the seafloor is created at ridges and consumed at trenches on a geologically short time scale. Other planets, notably Mars and Venus, have surface features that suggest some elements of plate tectonics, but none is known to be undergoing the constant rejuvenation of the surface as is the Earth.

The fundamental laws of geologic succession, of the differences between igneous rocks (those crystallized from a melt) and sedimentary rocks (those formed by diagenesis of sediments deposited by surface processes), began to be understood toward the end of the 18th century. At about the same time the measurement of the constant in Newton's law of gravitation showed that the specific gravity of the Earth was about 5.5, whereas that of a typical crustal rock was only about 2.7. Obviously, the interior must be much denser, and it became apparent that pressure alone could not explain the difference. Instead, there have to be differences in chemical composition, involving a decrease with depth of the light elements abundant in the crust (oxygen, silicon, and aluminum) and an increase of heavier elements such as iron. With reference again to the other planets of the solar system, it is seen that their densities tend to fall into two groups: the inner planets with densities close to that of the Earth, and the giant planets with appreciably lower densities. The members of the first group probably have iron cores (proposed for the Earth in the early years of the 20th century), while those of the second must contain large amounts of very light elements such as hydrogen.

If the Earth were completely static, it would be virtually impossible to obtain information about the interior apart from the mean density. Dynamism is, however, the mark of the Earth: tectonic plates move, probably driven by slow convection currents within the Earth's mantle; earthquakes occur; and, from the study of earthquake waves, the broad outlines of the interior can be established. The resulting picture of the Earth includes a solid inner core, a fluid outer core whose radius is more than half the planet's radius, a predominantly solid mantle, and a chemically distinct thin crust that contains most of the familiar geologic features. The rigid plates that are driven over the surface consist of the crust and some uppermost mantle, which together make up a mechanically distinct regime called the lithosphere. In addition to producing earthquakes, the motions of the Earth's interior bring to the surface rocks typical of the deep interior as well as heat. Indeed, scientists now realize that it is the heat escaping from the outer part of the Earth that drives the motions that shape the surface. Smaller bodies of the solar system (*e.g.*, Mars and the Moon) have probably cooled to such an extent that convectively driven tectonics no longer operate.

The heat now escaping from the Earth's surface probably comes in the main from radioactivity throughout the mantle. Some heat may still come from that generated when the dense core "fell" into the Earth's centre, and some may be original heat. The question of whether the Earth began hot or cold is not definitely settled, although majority opinion favours a cold origin with intense early heating through radioactivity and the separation of the metallic core. Ages determined by the analysis of radioactive isotopes and their daughter products provide a clue to early history. The Earth as a distinct body is known to have an age of 4.6×10^9 years, and samples of lunar material show a similar age. The oldest continental rocks currently found, however, in Canada, have ages (since the time they solidified) of approximately 3.9×10^9 years. Presumably the record at the Earth's surface of the first 700 million years was erased by the elevated temperatures that prevailed in those times.

That the Earth has a magnetic field has been known at least since the 11th century when the directional properties of suspended magnetic rock (magnetite; also called lodestone) were first used for navigation. Over the centuries the characteristics of this changing field have become better understood, until it now appears that the only plausible cause is some system of motions in the Earth's liquid outer core. These motions, which may be thermally driven like the slow convection currents in the mantle, constitute an electromagnetic dynamo whose electric currents sustain the field. Many problems remain: Why, for example, should the field have reversed polarity at irregular intervals through geologic time? But again, the existence of the field is in all likelihood simply further evidence of the Earth's dynamic structure.

Some other planets, notably Jupiter and Mercury, are known to have magnetic fields. It is interesting that Jupiter and Mercury differ appreciably in density and therefore composition, but both planets must have the internal motions necessary to constitute a dynamo. The Moon probably once had a magnetic field, but its internal dynamo apparently ceased long ago as the fluid interior froze.

The Earth's magnetic field shields the planet from the most direct effects of the ionized gas that constitutes the solar wind, carving out a cavity known as the magnetosphere. The existence of the magnetosphere has in all likelihood played a fundamental role in determining the nature of the Earth's atmosphere and its climate and therefore in the development of life. Yet, the verification of the magnetosphere's existence is a fairly recent accomplishment, dating only from the International Geophysical Year of 1957–58. (G.D.G.)

For coverage of related topics in the *Macropædia* and *Micropædia*, see the *Propædia*, sections 133, 211, 212, 213, and 231, and the *Index*.

This article is divided into the following sections:

The figure and dimensions of the Earth 570
  Early conceptions 570
  Determination of the Earth's figure: a historical review 570
    Spherical era
    Ellipsoidal era
    The concept of the geoid
  Earth dimensions—diameter, mass, density 575
The gravitational field of the Earth 575
  The nature of gravity 575
  Basic characteristics of the terrestrial field 576
    Variation with latitude
    Variation with elevation
    Variation with internal density distribution
  Measurement of gravitational acceleration 577
    Absolute measurements
    Relative measurements
  Interpretation of gravity data 579
    Isostasy
    The global significance of gravity anomalies
The magnetic field of the Earth 581
  Observations of the Earth's magnetic field 582
    Representation of the field
    Measurement of the field
  Characteristics of the Earth's magnetic field 583
  Sources of the steady magnetic field 585
    The geomagnetic dynamo
    Crustal magnetization
    The ionospheric dynamo
    The ring current
    The magnetopause current
    The magnetotail current
    Field-aligned currents
    Convective electrojets
  Sources of variation in the steady magnetic field 589
    Secular variation of the main field
    Reversals of the main field
    Variations in the ionospheric dynamo current
    Magnetic storms—growth of the ring current
    Magnetospheric substorms—unbalanced flux transfer
    Magnetohydrodynamic waves—magnetic pulsations
The structure and composition of the solid Earth 595
  Zonal structure as reflected by variations in physical properties 595
    Seismology: wave-velocity and density distributions
    Electrical conductivity
    Temperature distribution
    Rheological properties
  Zonal variations in chemical and mineralogical composition 601
  Development of the Earth's structure and composition 603
The major geologic features of the Earth's exterior 605
  Deformation of the crust 605
    Force, stress, and strain
    Geologic structures
  Physiographic expressions of crustal deformation 606
    Folding
    Faulting
    Formation of joints
  The surface of the Earth as a mosaic of plates 607
    Geometry and rates of plate movement
    Types of plate boundaries
    Activity along plate boundaries
    Intraplate activity
    The cause of plate motions
    Energy sources for convection
    Evidence for polar wandering, continental drift, and seafloor spreading
Bibliography 614

The figure and dimensions of the Earth

EARLY CONCEPTIONS

The definition of the figure of the Earth—*i.e.*, its size and shape—usually does not involve the description of mountains and valleys but, rather, the size and shape of the mean sea-level surface and its continuation under the land. This hypothetical surface, called a geoid, is a reference surface from which topographic heights and ocean depths are measured. Because of the irregular mass distributions in the Earth and the resultant gravity anomalies, the geoid is not a simple mathematical surface and consequently is not a suitable reference surface for a geometric figure of the Earth (see below).

As reference figures of the Earth, but not for its topography, simple geometric forms are used that approximate the geoid. For many purposes an adequate geometric representation of the Earth is a sphere, for which only the radius of the sphere must be stated. When a more accurate reference figure is required, an ellipsoid of revolution is used as a representation of the Earth's shape and size. It is a surface generated by rotating an ellipse 360° about its minor axis. An ellipsoid that is used in geodetic calculations to represent the Earth is called a reference ellipsoid.

Reference ellipsoid

This ellipsoid of revolution is the shape most often used to represent a simple geometric reference surface.

An ellipsoid of revolution is specified by two parameters: a semimajor axis (equatorial radius for the Earth) and a semiminor axis (polar radius), or the flattening. Flattening (f) is defined as the difference in magnitude between the semimajor axis (a) and the semiminor axis (b) divided by the semimajor axis, or $f = (a - b)/a$. For the Earth the semimajor axis and semiminor axis differ by about 21 kilometres, and the flattening is about one part in 300. The departures of the geoid from the best fitting ellipsoid of revolution are about ±100 metres (330 feet); the difference between the two semi-axes of the equatorial ellipse in the case of a triaxial ellipsoid fitting the Earth is only about 80 metres.

DETERMINATION OF THE EARTH'S FIGURE: A HISTORICAL REVIEW

Spherical era. *The ancients.* Credit for the idea that the Earth is spherical is usually given to Pythagoras (flourished 6th century BC) and his school, who reasoned that, because the Moon and the Sun are spherical, the Earth is too. Notable among other Greek philosophers, Hipparchus (2nd century BC) and Aristotle (4th century BC) came to the same conclusion. Aristotle devoted a part of his book *De caelo* (*On the Heavens*) to the defense of the doctrine. He also estimated that the circumference of the Earth is about 400,000 stadia (a Greek stadium varied in length locally from 154 to 215 metres). Since the length of his stadium is not known with certainty, the accuracy of his estimate cannot be established. This seems to be the first scientific attempt to estimate the size of the Earth. Eratosthenes (3rd century BC), however, is considered to be one of the founders of geodesy because he was the first to describe and apply a scientific measuring technique for determining the size of the Earth. He used a simple principle of estimating the size of a great circle passing through the North and South poles (Figure 1). Knowing the length of an arc (l) and the size of the corresponding central angle (α) that it subtends, one can obtain the radius of the sphere from the simple proportion that length of arc to size of the great circle (or circumference, $2\pi R$, in which R is the Earth's radius) equals central angle to the angle subtended by the whole circumference (360°):

Eratosthenes' measurement

$$l : 2\pi R = \alpha^\circ : 360^\circ. \qquad (1)$$

In order to determine the central angle α, Eratosthenes selected the city of Syene (modern Aswān on the Nile)

because there the Sun in midsummer shone at noon vertically into a well (Figure 1). He assumed that all sunrays reaching the Earth were parallel to one another, and he observed that the sunrays at Alexandria at the same time (midsummer at noontime) were not vertical but lay at an angle $^1/_{50}$ of a complete revolution of the Earth away from the zenith. Probably using data obtained by surveyors (official pacers), he estimated the distance (l) between Alexandria and Syene to be 5,000 stadia. From the above equation Eratosthenes obtained, for the length of a great circle, $50 \times 5{,}000 = 250{,}000$ stadia, which, using a plausible contemporary value for the stadium (185 metres), is 46,250,000 metres. The result is about 15 percent too large in comparison to modern measurements, but his result was extremely good considering the assumptions and the equipment with which the observations were made.

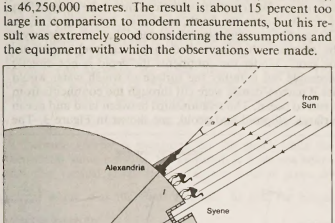


Figure 1: Eratosthenes' arc measuring method (see text).

The introduction of triangulation. A new era in determining the size of the Earth began through the introduction of triangulation. The idea of triangulation was apparently conceived by the Danish astronomer Tycho Brahe before the end of the 16th century, but it was developed as a science by a contemporary Dutch mathematician, Willebrord van Roijen Snell. Snell used a chain of 33 triangles to determine the length of an arc essentially in the way customarily done today. The resulting size of the Earth, however, was 3.4 percent too small. The idea of triangulation is to establish a network of stations that form connecting triangles. One side of the first triangle in the chain, called the baseline, and all angles of the triangles are accurately measured. Using the law of sines from spherical trigonometry, the lengths of all sides thus can be computed starting from a known baseline. When the lengths and angles are known, coordinates can be computed for each point provided that the coordinates of one point and one azimuth are known. Triangulation points are usually placed on the tops of the hills because the neighbouring points must be clearly visible. Commonly, more complicated figures such as quadrilaterals with diagonals are used in triangulation.

Snell's contribution

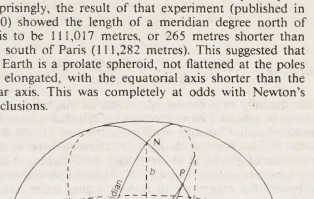
In 1669 Jean Picard, a French astronomer, first used a telescope in determining latitude and in measuring angles in triangulation that consisted of 13 triangles and extended from Paris 1.2° northward. His observations and results were extremely important because his length of arc on a great circle corresponding to 1° was used by the English physicist and mathematician Sir Isaac Newton in his theoretical calculations to prove that the attraction of the Earth is the principal force governing the motion of the Moon in its orbit.

Ellipsoidal era. The period from Eratosthenes to Picard can be called the spherical era of geodesy. A new ellipsoidal era was begun by Newton and the Dutch mathematician and scientist Christiaan Huygens. In Ptolemaic astronomy it had seemed natural to assume that the Earth was an exact sphere with a centre that, in turn, all too easily became regarded as the centre of the entire universe. However, with growing conviction that the Copernican system is true—the Earth moves around the Sun and rotates about its own axis—and with the advance in mechanical knowledge due chiefly to Newton and Huygens, it seemed natural to conceive of the Earth as an oblate spheroid. In one of the many brilliant analyses in his *Principia*, published in 1687, Newton deduced the Earth's shape theoretically and found that the equatorial semi-axis would be $^1/_{230}$ longer than the polar semi-axis (true value about $^1/_{300}$).

First estimate of flattening

Experimental evidence supporting this idea emerged in 1672 as the result of a French expedition to Guiana. The members of the expedition found that a pendulum clock that kept accurate time in Paris lost $2^1/_2$ minutes a day at Cayenne near the Equator. At that time no one knew how to interpret the observation, but Newton's theory that gravity must be stronger at the poles (because of closer proximity to the Earth's centre) than at the Equator was a logical explanation.

It is possible to determine whether or not the Earth is an oblate spheroid by measuring the length of an arc corresponding to a geodetic latitude difference at two places along the meridian (the ellipse passing through the poles) at different latitudes, which means at different distances from the Equator. This can be seen from Figure 2, in which the geodetic latitude at any point (P) is represented by the angle made between a line perpendicular to the ellipsoidal surface at the point P and the equatorial plane. This angle differs from the geocentric latitude that is determined by a line directed from the point P toward the Earth's centre. Such measurements of arc were made by the astronomer Gian Domenico Cassini and his son Jacques Cassini in France by continuing the arc of Picard north to Dunkirk and south to the boundary of Spain. Surprisingly, the result of that experiment (published in 1720) showed the length of a meridian degree north of Paris to be 111,017 metres, or 265 metres shorter than one south of Paris (111,282 metres). This suggested that the Earth is a prolate spheroid, not flattened at the poles but elongated, with the equatorial axis shorter than the polar axis. This was completely at odds with Newton's conclusions.



$OE = a$ = semi-major axis
$ON = b$ = semi-minor axis
$\angle PFA = \varphi$ = geodetic latitude
$\angle POA = \psi$ = geocentric latitude
$\angle BOA = \lambda$ = geodetic longitude
$OP = r$ = radius vector
N and S = poles
FP normal to ellipsoid at P

Figure 2: Elements of a reference ellipsoid.

In order to settle the controversy caused by Newton's theoretical derivations and the measurements of Cassini, the French Academy of Sciences sent two expeditions, one to Peru led by Pierre Bouguer and Charles-Marie de La Condamine to measure the length of a meridian degree in 1735 and another to Lapland in 1736 under Pierre-Louis Moreau de Maupertuis to make similar measurements. Both parties determined the length of the arcs using the method of triangulation. Only one baseline, 14.3 kilometres long, was measured in Lapland, and two baselines, 12.2 and 10.3 kilometres long, were used in Peru. Astronomic observations for latitude determinations from which the size of the angles was computed were made us-

Efforts to measure the length of a meridian degree

ing the zenith sectors having radii up to four metres. The expedition to Lapland returned in 1737, and Maupertuis reported that the length of one degree of the meridian in Lapland was 57,437.9 toises. (The toise was an old unit of length equal to 1.949 metres.) This result, when compared to the corresponding value of 57,060 toises near Paris, proved that the Earth was flattened at the poles. Later, large errors were found in the measurements, but they were in the "right direction."

After the expedition returned from Peru in 1743, Bouguer and La Condamine could not agree on one common interpretation of the observations, mainly because of the use of two baselines and the lack of suitable computing techniques. The mean values of the two lengths calculated by them gave the length of the degree as 56,753 toises, which confirmed the earlier finding of the flattening of the Earth. As a combined result of both expeditions, these values have been reported in the literature: semimajor axis, $a = 6{,}397{,}300$ metres; flattening, $f = 1/216.8$.

Almost simultaneously with the observations in South America, the French mathematical physicist Alexis-Claude Clairaut deduced the relationship between the variation in gravity between the Equator and the poles and the flattening. Clairaut's ideal Earth contained no lateral variations in density and was covered by an ocean, so that the external shape was an equipotential of its own attraction and rotational acceleration. Under these assumptions, gravity at sea level can be written as a function of latitude φ in the form

$$\gamma_s(\varphi) = \gamma_{\text{Equator}} [1 + B \sin^2 \varphi]. \quad (2)$$

The expression deduced by Clairaut is

$$B = \frac{5}{2} m - f, \quad (3)$$

where $m = \dfrac{\text{centrifugal acceleration at Equator}}{\text{attraction at Equator}}$.

The quantity m is on the same order of magnitude as f; it can be obtained more precisely by calculation than by measurement. Clairaut's result is accurate only to the first order in f, but it shows clearly the relationship between the variation of gravity at sea level and the flattening. Later workers, particularly Friedrich R. Helmert of Germany, extended the expression to include higher order terms, and gravimetric methods of determining f continued to be used, along with arc methods, up to the time when Earth-orbiting satellites were employed to make precise measurements (Table 1).

Significance of Clairaut's work

Table 1: Historical Determinations of the Earth's Radius and Flattening

author	year	method	equatorial radius (in metres)	1/f*
P. Bouguer and P.-L. M. de Maupertuis	1735–43	arc	6,397,300	216.80
G.B. Airy	1830	arc	6,376,542	299.30
A.R. Clarke	1866	arc	6,378,206	295.00
F.R. Helmert	1884	gravimetric		299.25
J.F. Hayford	1906	arc	6,378,283	297.80
W.A. Heiskanen	1928	gravimetric		297.00
H. Jeffreys	1948	arc	6,378,099	297.10

*Flattening denoted by f.

Numerous arc measurements were subsequently made, one of which was the historic French measurement used for definition of a unit of length. In 1791 the French National Assembly adopted the new length unit, called the metre and defined as 1:10,000,000 part of the meridian quadrant from the Equator to the pole along the meridian that runs through Paris. For this purpose a new and more accurate arc measurement was carried out between Dunkirk and Barcelona in 1792–98. These measurements combined with those from the Peruvian expedition yielded a value of 6,376,428 metres for the semimajor axis and $1/311.5$ for the flattening, which made the metre 0.02 percent "too short" from the intended definition.

Adoption of the metre unit

The length of the semimajor axis, a, and flattening, f, continued to be determined by the arc method but with modification for the next 200 years. Gradually instruments and methods improved, and the results became more accurate. Interpretation was made easier through introduction of the statistical method of least squares.

(U.A.U./G.D.G.)

The concept of the geoid. As noted above, the actual sea-level surface of the Earth, even in the absence of the effects of waves, winds, currents, and tides, is not a simple mathematical form. The unperturbed ocean surface must be an equipotential surface of the gravitational field, and because the latter reflects variations due to heterogeneities of density within the Earth, so also do the equipotentials. The particular equipotential surface that coincides over the oceans with unperturbed mean sea level constitutes the geoid. Under the continents the geoid is not directly accessible but is rather the surface to which water would rise if narrow canals were cut through the continents from ocean to ocean. The relationships between land and ocean surfaces, ellipsoid and geoid, are shown in Figure 3. The

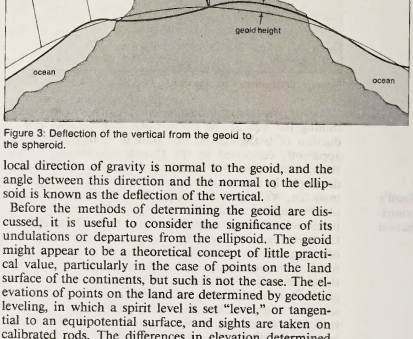


Figure 3: Deflection of the vertical from the geoid to the spheroid.

local direction of gravity is normal to the geoid, and the angle between this direction and the normal to the ellipsoid is known as the deflection of the vertical.

Before the methods of determining the geoid are discussed, it is useful to consider the significance of its undulations or departures from the ellipsoid. The geoid might appear to be a theoretical concept of little practical value, particularly in the case of points on the land surface of the continents, but such is not the case. The elevations of points on the land are determined by geodetic leveling, in which a spirit level is set "level," or tangential to an equipotential surface, and sights are taken on calibrated rods. The differences in elevation determined are therefore with respect to the equipotential and so very nearly with respect to the geoid. The determination in three coordinates of a point on the continental surface by classical techniques thus required the knowledge of four quantities: latitude, longitude, elevation above the geoid, and undulation of the geoid from the ellipsoid at that location. Furthermore, the deflection of the vertical played a most important role, since its components in orthogonal directions contributed errors of the same amounts in astronomical determinations of latitude and longitude. While geodetic triangulation provided relative horizontal positions with high accuracy, the networks of triangulation in each nation or continent began from points whose astronomical positions were assumed. The only possibility of connecting these networks into a global system lay in the computation of the deflections (*i.e.*, the slopes of the geoid) at all initial points. It is true that modern methods of geodetic positioning (discussed below) have altered this approach, but the geoid remains an important concept with definite practical utility.

Determining the form of the geoid with Stokes's formula. The geoid is in essence an equipotential surface of the actual gravitational field. In the vicinity of a local mass excess that adds potential ΔU to the normal Earth's potential at a point, the surface must warp outward in

order to keep the total potential constant. The undulation N is given by

$$N = \frac{\Delta U}{g}, \tag{4}$$

where g is the local value of the acceleration due to gravity. The effect of mass above the geoid complicates the simple picture; it can be allowed for in practice, but it is convenient to consider a point at sea level. The first problem is to determine N, not in terms of ΔU, which is not measured in terrestrial surveys, but rather in terms of departures of g from normal. The difference between the local measured value of gravity and the theoretical value at the same latitude on an ellipsoidal Earth free of lateral density variations is Δg. (The definition of Δg for points on the land surface above sea level is considered below.) The anomaly Δg arises from two causes. The first is the attraction of the mass excess, whose effect on gravity is given by the negative radial derivative of ΔU—*i.e.*, $-\frac{\partial}{\partial r}(\Delta U)$. The second is the effect of the height N, because gravity is measured on the geoid while the theoretical value refers to the ellipsoid. It is shown below that the vertical gradient of g at sea level is given by $\left(\frac{-2g}{a}\right)$ where a is the Earth's radius, so that the height effect is given by

$$\left(\frac{-2g}{a}\right) N = -\frac{2\,\Delta U}{a}. \tag{5}$$

Combining both effects, therefore,

$$\Delta g = -\frac{\partial}{\partial r}(\Delta U) - \frac{2\,\Delta U}{a}. \tag{6}$$

Formally, equation (6) establishes the relation between ΔU and the measurable value Δg, and if ΔU were determined, equation (4) would yield N. However, since both Δg and ΔU contain the effects of mass anomalies throughout an ill-defined region of the Earth, not just beneath the station, equation (4) cannot be solved at a point on the Earth without reference to others. The problem of relating N to Δg in a calculable manner was solved by the British physicist and mathematician Sir George Gabriel Stokes in 1849. Stokes obtained an integral equation for N, in which the integrand contains values of Δg, convolved with a function of their angular distance from the station, and the integral extends over the surface of the Earth. Until the launching of satellites in 1957, Stokes's formula constituted the principal method of determining the form of the geoid, but its application presented great difficulties. The function of angular distance contained in the integrand converges very slowly with that distance, and in the attempt to calculate N at any point—even in countries where g has been extensively measured—uncertainties enter from unsurveyed regions of the Earth that may be at considerable distances from the station. Various methods of extrapolating the gravity anomalies into these regions on the assumption of isostatic equilibrium (see below) were attempted, but the modern approach, which is to combine data from satellites and from ground observers, makes use of the expansion of the potential in spherical harmonic rather than Stokes's integral.

Problems associated with Stokes's formula

The contribution of orbiting satellites. The development of artificial satellites whose orbits could be observed from Earth totally revolutionized man's ability to define the shape of the Earth and its gravity field. A value for the flattening of the ellipsoid that superseded all previous values was obtained within weeks after the launching of the Soviet Sputnik I in 1957. Since that time, scientists have repeatedly refined the geoid with observations from a succession of Earth-orbiting satellites.

Refinement of the geoid

As a satellite moves through the Earth's gravitational field, it experiences forces, in addition to the central attraction, because of irregularities in that field. These forces perturb the orbit of the satellite from the simple form given by Kepler's laws of planetary motion.

It is usual to start with an expression for the potential U of the Earth's gravitational field, in spherical coordinates (r,θ,λ), with origin at the mass centre of the Earth. The gravitational potential U satisfies Laplace's equation, a widely used second-order partial differential equation named after the 18th-century French mathematician and astronomer Pierre-Simon Laplace. Accordingly, it can be expressed as a sum of spherical harmonics (a series of terms by which a variation of a quantity over the surface of a spherical or nearly spherical body such as the Earth can be expressed mathematically to any desired degree of accuracy):

$$U = \frac{MG}{r}\left[1 - \sum_{l=2}^{\infty} J_{l,0}\left(\frac{a}{r}\right)^l P_l(\cos\theta) + \sum_{l=2}^{\infty}\left(\frac{a}{r}\right)^l \sum_{m=1}^{l} J_{l,m} P_l^m(\cos\theta)\cos m(\lambda - \lambda_{lm})\right], \tag{7}$$

where M = the mass of the Earth; G = the gravitational constant; a = the equatorial radius of the Earth; θ = colatitude; and λ = longitude measured from an arbitrary meridian. The functions $P_l(\cos\theta)$ and $P_l^m(\cos\theta)$ are Legendre polynomials and Legendre associated polynomials (particular solutions of Laplace's equation), respectively. The quantities $J_{l,0}$ and $J_{l,m}$ are dimensionless numbers whose magnitudes give the relative importance of the different spherical harmonic terms (or "wavelengths") in the potential field. They were so designated in honour of Sir Harold Jeffreys, a pioneer in the analysis of the gravitational field in the pre-satellite era. Two features of equation (7) are important. If all of the J's were zero, U would have spherical symmetry and a satellite would move in a constant elliptical orbit, as deduced by Kepler. Properties of this orbit would yield a value for the product MG, but not for M or G separately. Similarly, all observations on actual orbits give the product MG (see below); the mass of the Earth can be determined only when G is measured independently. Second, equation (7) contains no terms in $l = 1$; this is a consequence of the selection of the mass centre of the Earth as origin, with all first moments of mass about that origin vanishing.

The most important term in the summations is that involving $J_{2,0}$. Inserting the value of $P_2(\cos\theta)$, the contribution to the potential is seen to be

$$\frac{3}{2}\frac{MG\,a^2}{r^3} J_{2,0}\left(\frac{1}{3} - \cos^2\theta\right).$$

The derivative of this expression with respect to θ is the force per unit mass acting on the satellite in the direction of increasing θ.

Physically, the term represents the effect on the potential of the ellipsoidal shape of the Earth, and it is not surprising, therefore, that $J_{2,0}$, known as the dynamical form factor, is closely related to the flattening f. In fact,

$$f = \frac{3}{2} J_{2,0} + \frac{m}{2}, \tag{8}$$

where m is the quantity introduced in equation (3).

As a satellite in an inclined orbit passes over the equatorial region of the Earth, it experiences a force toward the Equator as a result of the mass in the equatorial bulge. This force represents a torque about the origin, and as in the case of the spinning top or gyroscope, the application of the torque causes the rotation axis of the satellite (normal to the plane of the orbit) to precess about the Earth's rotation axis. The plane of the orbit therefore precesses, resulting in changes in the satellite's path that can be observed from Earth with a high degree of accuracy.

Analysis of the dynamics gives the angular velocity of precession, ω, as

$$\omega = \frac{3}{2} J_{2,0}\left(\frac{a}{r}\right)^2 \sqrt{g_r}\,\frac{a}{r^{1.5}}\cos i, \tag{9}$$

where g_r is the value of g at satellite height and i is the inclination of the orbit. The numerical value of $J_{2,0}$ is approximately 0.001; for a satellite at the height of roughly

740 kilometres, with $i = 20°$, equation (9) gives ω as 6.5° per day. Since the precession persists over the life of the satellite, the rate can be observed with great accuracy.

The higher degree zonal spherical harmonics (the first summation) in equation (7) lead to perturbations of the orbit in the precessing orbital plane. To achieve high accuracy in satellite tracking, special satellites carrying reflectors have been employed in conjunction with ground stations equipped with lasers that may be beamed at such satellites. The time that it takes for a laser pulse to travel to a satellite and back gives its instantaneous distance from the station. This technique represents an advance over an earlier method of geometric satellite geodesy in which a satellite was photographed simultaneously from a number of stations on Earth against a background of stars. That method did not require a precise knowledge of a satellite's orbit and permitted the location of an unknown station on Earth to be fixed relative to known stations. The simultaneous measurement of the distance from ground stations to satellites (*i.e.*, satellite trilateration) allows points on Earth to be precisely located when the orbit is well determined, but the latter, as indicated above, depends on a knowledge of the gravitational field being sought in the experiment. The solution to this apparent paradox is that sufficient observations be obtained to permit the optimum determination of both station coordinates in a global system and orbital parameter simultaneously.

A pioneer satellite designed for geodetic purposes was Lageos (Laser Geodynamic Satellite), launched by the United States on May 4, 1976, into a nearly circular orbit at a height of approximately 6,000 kilometres. It consisted of an aluminum sphere 60 centimetres in diameter that carried 426 reflectors suitable for reflecting laser beams back along their paths. The relatively high elevation was chosen to minimize both the effects of atmospheric drag and local gravity anomalies. Height, however, also attenuates the very effects that are sought, for, as equation (7) indicates, these decrease with increasing values of r. Satellites are therefore most effective at providing values of $J_{l,o}$ to about $l = 16$ (a wavelength on the order of 2,500 kilometres). For larger values of l, measurements of gravity on the Earth's surface, reduced to mean values representative of $1° \times 1°$ areas, must be used.

Satellite orbit analysis

The tesseral harmonics in equation (7), those terms involving $J_{l,m}$, present an additional difficulty. Because the terms represent contributions to the potential having a longitude dependence, their effect in general is averaged out as the Earth rotates under a satellite. The only exception is if the orbit and period of the satellite are such that the satellite tracks over points on the Earth equally separated in longitude, repeating each track precisely after an integral number of cycles. Such a satellite is said to be resonant to a particular value of m. Values of m for which resonant satellites can be found are limited, lying between 9 and 15. For other tesseral harmonics, observations of surface gravity must again be used.

A great advantage of the spherical harmonic expansion is that there is a simple relationship between weighting functions in the geoid undulations, $N_{l,m}$, gravity anomalies, $\Delta g_{l,m}$, and the $J_{l,m}$ terms. It is

$$N_{l,m} = \frac{a}{(l-1)\,g}\,(\Delta g_{l,m}) = a\,(J_{l,m} - J'_{l,m}). \qquad (10)$$

Equation (10) is an approximation to the extent that mean values of radius, a, and gravity, g, have been used. For the construction of maps of the geoid, however, it is usually sufficiently accurate, and it indicates clearly how the undulation coefficients $N_{l,m}$ can be obtained either from the gravity anomalies Δg or from the terms $J_{l,m}$ determined by satellites. The global map of the geoid is obtained by synthesizing the spherical harmonics, weighted by $N_{l,m}$, up to the maximum values of l and m available from the analysis of the observations.

Equation (10) is also useful for predicting the general nature of a map of the geoid, as contrasted to a map of gravity anomalies Δg. Each term in the expansion of N is reduced by the factor $1/(l-1)$ in comparison to the corresponding term in Δg. As l increases, the reduction becomes more significant in that local effects do not appear on the geoidal map.

Figure 4 shows a geoid determined from a combination of satellite observations, including Lageos, and surface measurements of gravity. The departures of the geoid from the ellipsoid range up to about 100 metres, the most pronounced inward warp lying just south of India. There is no obvious direct correlation between continents and oceans, but there are correlations with some of the major features of global tectonics.

Radar altimetry of the ocean surface. As noted above, the geoid over the oceans coincides with mean sea level provided that the dynamic effects of winds, tides, and currents are removed. The surface of the sea acts as a reflector for radar waves, and a satellite equipped with a radar altimeter can be used to sound from the satellite's instantaneous position to the sea. The accuracy with which the sea surface can be reconstructed depends on how precisely the satellite orbit is known, and the reduction of the dynamic effects on the sea surface (waves and semidiurnal and diurnal tides) depends on averaging—

After C. Reigber *et al., Journal of Geophysical Research*, no. 90, p. 9292 (1985); copyright American Geophysical Union

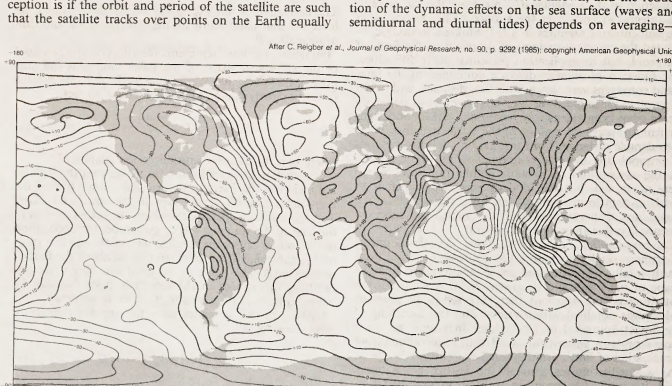


Figure 4: The geoid with respect to an ellipsoid with a flattening, f, of 1/298.257. The contours are at intervals of 10 metres.

over several days—of heights obtained from successive passes over identical points on the Earth.

Mapping of the sea surface with Seasat

The first satellite dedicated to mapping the ocean surface was Seasat 1, launched by the United States on June 26, 1978. Seasat was operational until Oct. 10, 1978, and reproduced its path over the Earth every three days. It sampled elevation every three kilometres along the track and thereby provided average ocean heights for literally millions of points on the sea surface. The precision of a single determination of satellite height above the ocean surface was a few centimetres.

A global map produced from 18-day averages of Seasat elevations is shown in Figure 5. While this is not strictly the geoid because long-term dynamic effects such as those of currents have not been averaged out, it is very close to it. Comparisons between the Seasat map and the geoids determined by the method described above show agreement to about one metre, which is estimated to be the maximum dynamic effect on sea surface "topography." The differences between true geoidal maps and maps of the sea surface are expected eventually to form a powerful tool for physical oceanography. Thus far, the main contribution of Seasat has been to provide a direct visual confirmation of the reality of the oceanic geoid and observations of higher resolution of some parts of the world ocean, as evidenced by a comparison of Figures 4 and 5. The tectonic significance of some of the major features of Figure 5 is discussed below.

EARTH DIMENSIONS—DIAMETER, MASS, DENSITY

As previously noted, terrestrial arc measurements are capable of yielding a value of the equatorial radius of the Earth, but satellite measurements are greatly superior for determining the flattening. After 10 years of satellite observations the International Union of Geodesy and Geophysics adopted the Geodetic Reference System 1967, defining $a_{equatorial}$, MG, and $J_{2,0}$. Minor revisions to the numerical values were made in 1983. The revised values are as follows:

Geodetic Reference System 1967

$$a_{equatorial} = (6{,}378{,}136 \pm 1)\ \text{m},$$
$$MG = (39{,}860{,}044 \pm 1) \times 10^{7}\ \text{m}^3/\text{s}^2,$$
$$M_a G = (35 \pm 0.3) \times 10^{7}\ \text{m}^3/\text{s}^2$$
where M_a = mass of the atmosphere,
$$J_{2,0} = (108{,}262.9 \pm 1) \times 10^{-8}.$$

The adopted value of $J_{2,0}$ corresponds to a flattening of $1/298.257$.

While satellite observations determine the value of the product MG to eight significant figures, they cannot determine M and G separately. Because satellites orbit (in general) above the atmosphere, the value of M includes the mass of the atmosphere, but, as shown above, the contribution of the latter to MG is extremely small. The gravitational constant G, measured in the laboratory, is known much less accurately; it is $G = 6.67259 \times 10^{-11}$ $\text{m}^3\text{s}^{-2}\ \text{kg}^{-1}$ (with uncertainty in the last place of decimals). The combination of the laboratory value of G and the adopted value of MG results in a value for the mass of the Earth (including the atmosphere) of $M = 5.98 \times 10^{24}$ kg. With the volume determined by $a_{equatorial}$, the flattening, and the portions above sea level, this value of the mass gives the mean density $\bar{\rho} = 5{,}517\ \text{kg/m}^3$.

There is some indication that $J_{2,0}$ varies slowly with time, as will be discussed below when gravity anomalies are considered in relation to Earth structure. There have been suggestions that G has varied with time throughout the history of the universe and that it is scale-dependent. In the latter case, values determined in the laboratory would not be appropriate for terrestrial or astronomical problems. Evidence for either a time- or scale-dependence of G remains inconclusive.

For many years there has been speculation about the extent to which the actual flattening of the ellipsoid coincides with the theoretical form of a mass of fluid, of the same mass and rotation rate as the Earth, in hydrostatic equilibrium under its own attraction and rotational acceleration. In the pre-satellite era, neither the actual flattening nor the theoretical form were known with sufficient accuracy to permit a meaningful comparison. Recent estimates of the flattening, in the case of hydrostatic equilibrium, for a body free of lateral density variations are close to $1/299.5$; the actual flattening, $1/298.257$, is therefore slightly greater. Although some investigators have suggested that the discrepancy represents an inheritance from the time when the Earth was rotating more rapidly on its axis, the most probable explanation is that it is simply one effect of the recognized lateral heterogeneity in density of the real Earth.

The gravitational field of the Earth

THE NATURE OF GRAVITY

It is a familiar phenomenon that an object released above the Earth's surface accelerates toward the Earth. This phenomenon is a special case of universal gravitation—all mass within the universe attracts all other mass. The acceleration in this special case is known as the acceleration due to gravity, denoted g. Reference has already been made above to the fact that g varies over the Earth's surface and that this variation is intimately related to the shape of the sea-level surface, or geoid. In this section the nature of gravity, its measurement, and the relationship of gravity variations to the internal structure of the Earth are discussed.

Acceleration due to gravity

After J.G. March and T.V. Martin, *Journal of Geophysical Research*, no. 87, p. 3276 (1982); copyright American Geophysical Union

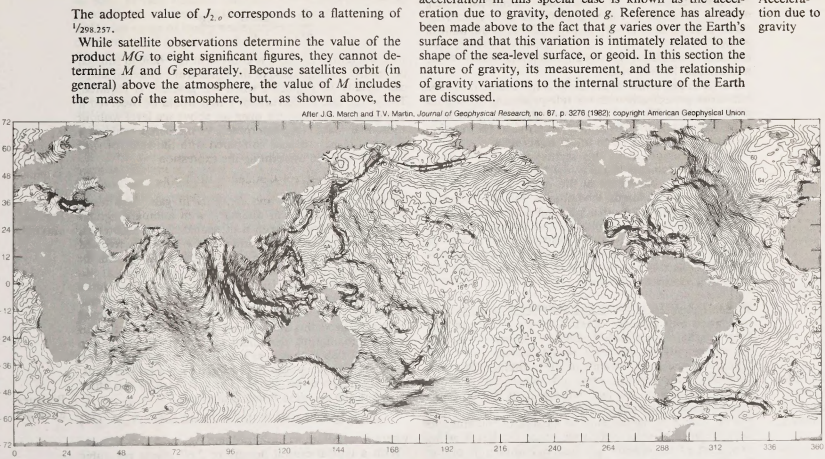


Figure 5: Sea-surface topography based on altimeter data from the Seasat 1 Earth-orbiting satellite. The elevation contours of the mean sea surface are at two-metre intervals.

Newton put forth the law of gravitation for particles of mass m_1 and m_2 separated by a distance r in the form

$$F_{12} = \frac{Gm_1m_2}{r^2}, \tag{11}$$

where F_{12} is the force of attraction of either particle on the other and G is a constant whose numerical value depends on the system of units employed. The application of Newton's law to bodies rather than to particles involves, in general, integration of the effects between differential elements; however, Newton also established the very convenient result that a uniform sphere attracts as though its mass were concentrated at the centre. This led to the possibility of measuring G in the laboratory, first exploited by the English physicist and chemist Henry Cavendish, by observing the force between massive spheres.

In the International System of Units (SI), the modern (as revised in 1986) value for the gravitational constant is $G = 6.67259 \times 10^{-11}$ m^3s^{-2} kg^{-1}. While the numerical value depends on the system of units, the apparently small magnitude of G is real. Gravitational forces between bodies of less than terrestrial size are indeed small, as compared, for example, to electrostatic forces or the magnetic forces between electric currents.

The acceleration due to gravity, g, is the force on a unit mass. In equation (11), if $m_1 = 1$ and $m_2 = M$, the mass of the Earth, the value of g on a spherical, uniform, non-rotating Earth is found to be

$$g = \frac{GM}{a^2}, \tag{12}$$

where a is the Earth's radius.

On the real Earth, departures from the spherical shape and uniformity and the effect of rotation all cause g to vary over the planet's surface. For example, the average value of g is close to 980 centimetres per second per second, but values at sea level range from about 978 near the Equator to 983 at the poles. Superimposed on this variation are the effects of internal structure, usually a small part of one centimetre per second per second. To describe these variations, a smaller unit has been introduced. In geophysics, one centimetre per second per second is known as the gal (after Galileo); 1×10^{-3} gals equal one milligal (mgal), which is the usual unit in which internal effects are measured. Since g itself is very nearly 1,000 gals, the milligal is approximately one part in 1,000,000 of g itself. Modern methods of measuring gravity approach a precision of one microgal, or 1×10^{-3} mgal.

Gravity at any point on the Earth is not constant in time but varies periodically with the tide-producing attractions of the Sun and Moon. The tidal variation has been measured for some years, and its analysis provides information on the yielding of the Earth under tidal forces.

While the emphasis in this section is on the uses of gravity measurements to study the Earth itself, it should be noted that a knowledge of the value of g is required, particularly in standards laboratories, for the measurement or calibration of other physical quantities. Such is the case whenever the weight of a known mass m is used as a standard of force, as, for example, in the absolute pressure exerted by a column of mercury in a barometer. The geophysicist is usually called upon to provide the best value of g for these locations.

BASIC CHARACTERISTICS OF THE TERRESTRIAL FIELD

Variation with latitude. Even if the Earth were of uniform density or uniformly stratified in layers of constant density, gravity at sea level would increase from the Equator to the poles because of the combined effects of the planet's rotation and spheroidal shape. The effect of rotation arises from the fact that any body on the Earth experiences a centripetal acceleration given by $r\omega^2$, where r is the perpendicular distance to the axis of rotation and ω is the angular velocity of rotation of the Earth on its axis. Part of the inward gravitational attraction of the Earth must provide centripetal acceleration simply to hold the body on the planet's surface and thus does not appear in the weight of the body or in measured g. Gravity therefore decreases as r increases from the poles to the Equator. Rotation, however, also distorts the sea-level shape into the ellipsoidal form, so that points near the poles are closer to the Earth's centre of attraction. The effects of rotation and shape are thus cumulative.

The effect of the Earth's rotation

As mentioned earlier, the variation of gravity on this ideal ellipsoidal Earth was investigated by Clairaut, who showed that the expected relationship was

$$\gamma_o(\varphi) = \gamma_{\text{Equator}}\,[1 + B \sin^2 \varphi], \tag{13}$$

where $\gamma_o(\varphi)$ is the sea-level value of gravity at latitude φ and B is a constant incorporating the effects of shape and rotation. If the potential of this field is expanded as a series of spherical harmonics (see above), it is found to contain a single latitude-dependent term involving $P_2(\cos\theta)$ where θ is colatitude. The numerical coefficient of this term is known as $J_{2,o}$. The three quantities B, $J_{2,o}$, and the flattening f of the ellipsoid are obviously interrelated, the relationships having been given above in equations (3) and (8).

Clairaut's equation as given is accurate only to the order of f; analysis to higher order of small quantities shows that additional terms in $(\sin \varphi)$ are involved. The quantity γ_o is a fundamental reference against which measured values of g may be compared to study all effects other than those of latitude. In the pre-satellite era the constants in the expression were obtained by fitting the measured values of g, distributed over the Earth, to the theoretical form; adoption of the constants by the International Union of Geodesy and Geophysics led to the International Gravity Formula. With the international adoption of MG, a, and $J_{2,o}$, it is more precise to compute γ_o. The adopted values of the latter lead to

$$\gamma_o = 978.03185\,[1 + 0.005278895 \sin^2\varphi + 0.000023462 \sin^4 \varphi] \text{ gals}. \tag{14}$$

Variation with elevation. Gravity on the continents of the Earth is rarely measured at sea level, so that, if the measured value of g is to be compared with γ_o a correction for height must be applied.

From the expression for gravity on a spherical Earth, the effect of increasing the distance from the Earth's centre is immediately obtained, as

$$\frac{\partial g}{\partial r} = -\frac{2GM}{r^3} = \frac{-2g}{r}. \tag{15}$$

At sea level the gradient is equivalent to -0.3086 milligal per metre; the departure of the Earth from spherical shape can usually be neglected when the gradient is required. This is the rate of decrease of g that would be observed if one went up through the open air above sea level without additional mass being interposed. A measured value of g at height h may then be compared with the value of γ_o at the same latitude by forming the expression

$$\Delta g_{\text{F.A.}} = (g + 0.3086h \times 10^{-3}) - \gamma_o \tag{16}$$

where h is in metres and $\Delta g_{\text{F.A.}}$ is in gals. The result is known as the free-air anomaly, with anomaly signifying the difference between a measured and a theoretical quantity and free-air indicating that g has been reduced to sea level by application of only the free-air term. In comparing points on the land surface with those at sea level, however, there is the additional effect at the higher point of the attraction of mass above sea level. Pierre Bouguer, whose work on arc measurements was noted earlier, realized this and applied a negative correction to g by approximating the actual topography to an infinite horizontal slab of thickness h and uniform density ρ. The attraction of such a slab, in centimetre-gram-second units, is easily shown to be $2\pi G\rho h$. The provision of this effect results in the so-called Bouguer anomaly

Free-air anomaly

Bouguer anomaly

$$\Delta g_B = [g + (0.3086 \times 10^{-3} - 200\pi G\rho)\, h] - \gamma_o \tag{17}$$

where h, as before, is in metres and Δg_B is in gals.

With a typical crustal density of 2.67 grams per cubic centimetre, the inclusion of the Bouguer term predicts that gravity decreases with increasing height of the land surface

at approximately 0.20 milligal per metre, as contrasted with the free-air decrease of 0.3086 milligal per metre. For gravity stations at sea, a slightly different Bouguer term is employed, which effectively replaces the water under the station with rock of normal crustal density.

At first sight, Bouguer's approximation for mass above sea level appears very crude, particularly in regions of irregular topography. In practice, however, it has turned out to be extremely efficacious. In most cases, an additional correction for the terrain in the immediate vicinity of the station only has to be applied within mountain belts.

Free-air anomalies and Bouguer anomalies both find a place in geodesy and geophysics, but it is important to recognize their difference. Because no allowance is made for the effect of mass above sea level in computing the free-air anomaly, the full effect of this mass remains. Indeed, the free-air anomaly is very nearly what would be observed if all of the continents were condensed to sheets of mass "buttered" over the geoid, allowing gravity to be measured everywhere on that surface. The free-air anomaly is therefore used in geoidal computations; it is, for example, the quantity Δg that was introduced earlier (equations 6 and 10). Nevertheless, as the attraction of the topography is not provided for, the free-air anomaly is strongly correlated locally with height to the extent that effects internal to the Earth are obscured. For this reason, Bouguer anomalies are normally used for the study of internal density variations. Yet again, on a broader scale, the Bouguer anomaly itself shows strong correlations with height, rising to positive values of several hundred milligals over the oceans and plunging to negative values of the same magnitude over mountain ranges. In extreme cases, the free-air anomalies remain closer to zero than do the Bouguer anomalies, suggesting that the Bouguer term has greatly overcorrected for the attraction of the topography. Such can only be the case if topographically high regions are somehow compensated for by mass deficiencies, or "roots," beneath and the oceans by rocks of greater than normal density beneath the seafloor (sometimes called "anti-roots"). This is the concept of isostatic compensation (see below).

Variation with internal density distribution. If the Earth were horizontally stratified so that density varied only with depth, the Bouguer anomaly would be everywhere close to zero. In fact, because the Earth is heterogeneous, the gravitational anomaly field displays "roughness" over a very wide range of length scales and amplitudes. When the distribution of excess or deficiency in density is known, the effect on gravity at any point can be determined by the straightforward application of Newton's law (equation 15) to all elements, making it possible to obtain the vertical component and integration over the limits of the distribution.

Geophysicists seek, however, to infer density distributions and derive from them insights into geologic structures. This process is known as interpretation; it is nonunique because different structures can produce the same anomaly at the surface of the Earth. This fact is clearly apparent in the very local anomaly shown in Figure 6. Since a sphere attracts as though its mass were concentrated at the centre, all spheres centred at the same point and representing the same excess mass produce identical anomaly curves. The diagram indicates two possibilities: The larger sphere could be a mass of more basic rock, or the smaller sphere a mass of metallic ore. Without further information, however, the densities and radius cannot be separated. The shape of the profile does, nevertheless, indicate the depth to the centre.

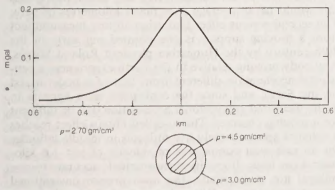


Figure 6: Two spherical dense bodies that produce identical Bouguer anomalies.

In general, gravitational effects increase with the scale of a structure. If the length scales in Figure 6 were increased by a factor of 100, yielding a structure of crustal dimensions, the peak anomaly would be 20 milligals rather than 0.2 milligal (though on this scale a body within the crust having a density of 4.5 grams per cubic centimetre would be most unlikely on geologic grounds).

The Earth contains a number of interfaces at which the normal variation of density with depth suffers discontinuities. These include the Mohorovičić discontinuity (or Moho) at the base of the crust, the core–mantle interface, and surfaces of mineralogical phase change within the mantle. The warping of any of these boundaries from its normal position can produce gravity anomalies at the Earth's surface. Since the amplitude of warp is normally very small compared with the depth of the interface, the effect can be represented by a surface distribution of anomalous mass smeared over the interface. If the densities above and below the interface are ρ_1 and ρ_2, an upward warp of amplitude y is equivalent to a surface density

Warping of the Earth's boundaries as a source of gravity anomalies

$$\sigma = (\rho_2 - \rho_1)\, y. \tag{18}$$

When the free-air gravity anomaly field Δg is available as the sum of spherical harmonic terms, these terms may be inverted directly to give the equivalent term in the spherical harmonic expansion for σ, provided the depth z of the interface is specified. The theory of spherical harmonics gives

$$\sigma_l = \frac{1}{4\pi G}\left(\frac{2l+1}{l+1}\right)\left(\frac{a}{a-z}\right)^{l+2}\Delta g_l, \tag{19}$$

where a is the Earth's radius and l is the degree of the spherical harmonic. The solution obviously becomes unstable at larger values of l in terms of the assigned z. In the expansion of the Earth's field, however, the first few terms beginning with $l = 2$ (but omitting the flattening term $J_{2,0}$) could be due to long wavelength warpings of boundaries as deep as the core–mantle interface. On the other hand, any warping of a surface of discontinuity in density represents a departure from hydrostatic equilibrium, which must be dynamically maintained by motions of the Earth material. Density variations through the material will produce gravitational effects that are intermingled with those of the boundary itself. There is, as yet, no final explanation for the lowest harmonics in the gravity field. The most promising option appears to be a search for correlations with broad anomalies in seismic wave velocity by means of seismic tomography (see below *Seismology: Wave-velocity and density distributions*).

MEASUREMENT OF GRAVITATIONAL ACCELERATION

For the study of the shape of the geoid or of large-scale structures in the Earth, values of g are required with a precision of one milligal or better. This implies a measurement accurate to one part in 10^6 of g itself. For more local studies, as in the use of gravity measurements for geophysical petroleum and mineral prospecting and for the study of the tidal variation of g at a fixed site, much higher precision approaching the microgal level is required. In order to provide for all requirements, it is convenient to separate measuring systems into two classes: (1) absolute systems, in which the actual value of g in correct acceleration units is obtained at a site without reference to any other point; and (2) relative systems, in which differences in g between stations or the time variation of g at one station are measured. To the accuracy sought, the first of these is much the more difficult.

Absolute measurements. The most fundamental and straightforward method of measuring g is to measure directly the acceleration of a body falling in a vacuum toward Earth, but until the mid-1900s timing systems of sufficient accuracy were simply not available. All older absolute measurements were based on measuring the period

Pendulum measurements

of swing of some form of physical pendulum. The period of a simple or mathematical pendulum is known to be

$$T = 2\pi\sqrt{l/g}, \tag{20}$$

where l is the length from the support to the hypothetical "bob" in which all mass is concentrated. The simple pendulum does not exist, but in any real form that can be swung from supports, two points, on either side of the mass centre, can always be found for which the periods are equal. The distance between these points is equal to l, the length of the equivalent simple pendulum, so that when equality of periods is found by experiment and adjustment, equation (20) can be solved for g. The reversible physical pendulum, as it is known, was employed by the British physicist Henry Kater in 1818 and by later workers up to 1940. Until 1968 all quoted values of g over the Earth were based on pendulum measurements made in Potsdam, Germany, in 1906, other points being determined by relative measurements from that site. By the middle of the century it became apparent that the Potsdam standard was in error by at least 15 milligals, but the methods then available were still not able to provide an absolute value with which to correct to one milligal. In 1952 Charles Volet of France described observations based on a free-fall apparatus in which a graduated rule was dropped in a vacuum in a tall chamber and photographed at precise time intervals as it passed the optic axis of a camera. Alignment of the graduations with a reference line on the photographs permitted g to be determined through the elementary equation

$$s = \frac{1}{2} gt^2, \tag{21}$$

where s is the distance fallen from rest in time t.

Variations and modifications of the falling-rule type of apparatus were employed in other countries. In 1967 A.H. Cook published the results of measurements made in London in which the object dropped was not a rule but a small sphere, which was recorded as it passed two optical systems. The small-body technique led the way to portable absolute systems, as opposed to fixed installations in standard laboratories. The first portable system was developed by James A. Hammond and James E. Faller in the United States in 1967 largely as a result of the availability of lasers that could produce highly directional beams of coherent light. In this system, the object dropped is a corner cube reflector, which reflects a light beam back along precisely the path through which it enters the evacuated chamber. Laser light directed into the chamber is split by a partial reflector, part of the beam traveling upward to be reflected by the falling cube and part of it being reflected in an identical fixed cube. The mixed light beam leaving the chamber displays alternating maxima and minima of intensity as the cube falls, depending on whether the mixing beams are in phase or antiphase. A photomultiplier tube measures the light. When the output of the photomultiplier is recorded on a precise time base, the falling cube can literally be tracked wavelength by wavelength of the laser light, and g can be determined by equation (21). A single drop can be made in less than one minute and an absolute value of g with an accuracy approaching 20 microgals obtained. The Hammond-Faller apparatus is not a field instrument, but it does provide a means for making precise absolute measurements at many points over the Earth.

Measurement with laser-based systems

Relative measurements. For studying Earth structures, instruments that will measure the differences in g from points where g is known are required. They have to be rugged enough to be taken into the field and capable of providing a reading in a matter of a few minutes. Instruments suitable for measuring g on ships at sea also are needed. As long ago as 1849, Sir John Herschel of England proposed an instrument consisting of a fixed mass m suspended on a spring. As the instrument was taken to different places, the varying weight mg of the mass produced differences in the extension of the spring. Yet, even for relative measurements, Herschel's instrument lacked the required sensitivity. Until about 1930 all relative gravity measurements as well as absolute measurements on the Earth were made with physical pendulums. When equation (20) is applied to two stations, 1 and 2, and l is the constant-equivalent simple pendulum length, the ratio of the values of gravity is immediately obtained in terms of the periods:

$$\frac{g_1}{g_2} = \frac{T_2^2}{T_1^2}. \tag{22}$$

Pendulum measurements, however, are necessarily very time consuming, and the realization that variations in g could be utilized in oil exploration led to intensive research on improving the static gravimeter. The difficulties Herschel had encountered were overcome by choosing suitably stable material for the springs, by employing mechanical, electrical, or optical devices to amplify the small displacements of the system, and by providing temperature control or compensation.

Such improvements are very much apparent in the Worden gravimeter, an example of a modern, highly portable system. Gravimeters of this type weigh only a few pounds and can measure differences in g to about 0.01 milligal in a few moments. All such instruments, however, have two limitations. As a system ages, readings at a single station will change over and above the true tidal change of g. This instrumental "drift" must be corrected by arranging observations in closed circuits so that readings at base points are repeated. Second, all differences in gravity are obtained initially in arbitrary scale divisions. Consequently, the instruments must be calibrated by measuring differences between points where g is known.

Gravity measurements from ships and aircraft

When measurements of gravity are made on a moving ship or in an aircraft, new effects come into play, and these require the most careful consideration if useful results are to be obtained. The most important of these effects is the fluctuating vertical acceleration of the vehicle relative to the Earth; its instantaneous value cannot be separated from an instantaneous measurement of g. On a ship, for example, vertical accelerations of thousands of milligals are possible, even in times of moderate swell. To measure g to a precision of one milligal at first appears impossible, but averaging over time makes it possible to reduce the effect. If the instantaneous vertical acceleration of the vehicle is $\frac{d^2z}{dt^2}$, its average over a time interval T is

$$\frac{1}{T}\int_0^T \left(\frac{d^2z}{dt^2}\right) dt = \frac{1}{T}\left[\left(\frac{dz}{dt}\right)_T - \left(\frac{dz}{dt}\right)_0\right]. \tag{23}$$

The quantity in brackets involves the vertical velocities of the vehicle at the beginning and end of the averaging interval. In some cases, this quantity can be measured using precise positioning techniques. Otherwise, the averaging interval T must be made sufficiently long for the quotient to be negligible. For ship-borne observations, recording gravimeters mounted on stabilized platforms to eliminate other disturbances are used in conjunction with computers to average the response over any adopted time interval T. The remaining problem, which is particularly serious in measurements from aircraft, is that there is a trade-off between accuracy and resolution. Only a single corrected value of g is obtained corresponding to the time interval T (which can be several minutes), and at aircraft speeds this would be an average over many kilometres of path. The second serious effect that arises in any measurement from a moving support is one pointed out early in the 20th century by the Hungarian physicist Roland Eötvös. Any body moving relative to the Earth experiences a centripetal acceleration different from that of a body at rest on the Earth, and since the centripetal term appears in g, the measured gravitational acceleration will be different for a moving observer. The so-called Eötvös effect reaches a value of approximately five milligals in middle latitudes for an east–west component of velocity of just 1.6 kilometres per hour. If gravity measurements accurate to one milligal are to be made, the most precise navigational systems are required. When these are available, shipborne surveys can be carried out with considerable success. The

airborne measurement of g remains experimental and appears to be limited to high-altitude surveys of a very general nature and low resolution. Some of the problems of measuring g from a moving vehicle are removed if, instead of g itself, the horizontal gradients of gravity are measured. Research is in progress on gradiometers that could be carried on aircraft or even on orbiting satellites.

For recording small tidal changes in gravity at a fixed site, a great advance has been made through the application of the low-temperature phenomenon of electrical superconductivity. In one gravimeter based on this phenomenon, a metallic sphere cooled by liquid helium to superconducting temperature is held in the constant magnetic field of superconducting coils. Changes in gravity tend to displace the sphere, but this is opposed by an electric field applied by capacitor plates above and below the sphere. The quantity recorded is the voltage that must be applied to these plates. Instruments of this kind measure variations in g to an accuracy approaching one microgal and have proved to be extremely valuable in the study of the tidal variations in gravity.

INTERPRETATION OF GRAVITY DATA

Isostasy. The general concept that higher portions of the Earth's crust are lighter and are maintained in position by correspondingly light roots beneath is very old. The quantitative description of this "compensation" of the Earth's topographic features dates from the geodetic triangulation of India in which the effect of the attraction of the Himalayas on the deflection of the vertical was accurately measured. In 1855 John Henry Pratt, an English amateur scientist, showed that the attraction was everywhere less than would be produced by the visible mass of the mountains. That same year the English astronomer George Biddell Airy offered the interpretation that the surface features of the Earth are formed in a crust whose density is less than that of the substratum; highly elevated regions of the Earth have crustal roots extending to greater depth (Figure 7). Pratt proposed an alternative model in which the density of crustal columns varies inversely with their height. In both cases, the weight of any crustal column of unit cross-sectional area exerted at the "level of compensation" is everywhere the same. Airy's description of his proposed mechanism was remarkable for the time because it preceded by more than 50 years the seismological identification of the crust and mantle (see below).

The Pratt and Airy proposals

From G.D. Garland, *Introduction to Geophysics*, 2nd ed., copyright © 1971 & 1979 by W.B. Saunders Co., reprinted by permission of Holt, Rinehart & Winston, Inc.

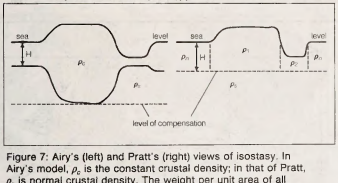


Figure 7: Airy's (left) and Pratt's (right) views of isostasy. In Airy's model, ρ_c is the constant crustal density; in that of Pratt, ρ_c is normal crustal density. The weight per unit area of all columns of the Earth is equal at the level of compensation.

The term isostasy was introduced in 1889 by the American geologist Clarence E. Dutton to describe the condition of equal pressure beneath a certain depth. During the early decades of the 20th century both the Pratt and Airy mechanisms were subjected to quantitative tests in which anomalies in the magnitude of gravity were first employed. Geodesists, led by John F. Hayford of the United States, tended to investigate the Pratt mechanism, while geophysicists, led by Weikko A. Heiskanen of Finland, concentrated on that of Airy. In both cases the approach was to compute new types of gravity anomalies known as isostatic anomalies, in which the effect of the roots on gravity, computed from the visible topography under the assumption of one or the other mechanism and model parameters (crustal thickness, densities), was added to the Bouguer anomaly. Model parameters were then varied until the closest approach to zero anomaly at all stations was obtained. In general, the best fit with the Pratt mechanism was achieved with a crustal thickness of about 100 kilometres, and with the Airy mechanism, a thickness of the crust for regions at sea level of 30–40 kilometres. The latter figures, in general agreement with the seismological thickness of the crust, supported the view that the Mohorovičić discontinuity (the interface between the crust and mantle) was the base of the Airy crust. With either model, there remain significant residuals that indicate local departures from the ideal isostatic condition.

As they stand, both the Pratt and Airy mechanisms are now known to be oversimplifications as compared with the real Earth. Yet, the ideas of Pratt and Hayford were to have an impact on the modern concept of plate tectonics (see below). The American geologist Joseph Barrell, analyzing their results, pointed out that the portion of the Earth above the level of compensation must be rigid (otherwise mountains would not stand up at all), and he dubbed this portion (apparently about 100 kilometres thick) the lithosphere and called the yielding region beneath the asthenosphere.

That some form of isostatic compensation is a fundamental property of the Earth can be seen in several ways: in the lack of expression of the continents and oceans in the global map of the geoid (Figure 4) and in the universal tendency of Bouguer anomalies to trend toward large negative values in elevated regions.

Isostatic compensation as a fundamental property

Modern investigations of isostasy on a statistical basis take as their starting point the correlation between Bouguer anomaly and height in regions of continental size. In such a region, both the Bouguer anomaly Δg and the topographic height h are Fourier transformed into functions of vector wave number k (*i.e.*, $|k| = 2\pi/\text{wavelength}$). Then, when there is correlation

$$\Delta g(k) = Q(k) \cdot h(k) + \text{residual}. \qquad (24)$$

Here $Q(k)$, measured in milligals per metres of height, is known as the isostatic response function, a term introduced by Leroy M. Dorman and Brian T.R. Lewis in a study of gravity over the continental United States undertaken during the early 1970s. A typical variation of $Q(k)$ with wavelength is shown in Figure 8. Since the Bouguer anomaly measures essentially the effect of the negative compensating masses, or roots, the form of the curve can be used to indicate their depth. For example, in the case of local Airy compensation, with the roots condensed onto a plane at depth z, the response function is

$$Q(k) = -2\pi\rho G \exp(-kz), \qquad (25)$$

where ρ is the crustal density.

The value of z may then be determined to best fit the observed response curve. The procedure may be generalized to cases where the compensation is distributed over depth and the response curve inverted to give the density deficiency as a function of depth. In most of the continental areas that have been analyzed, the compensating masses appear to be concentrated near the Mohorovičić discontinuity. This interface represents the main discontinuity in density in the outer part of the Earth (perhaps as much as 0.5 gram per cubic centimetre between lower crustal and mantle material). By contrast, the base of the lithosphere, which is marked by a change in mechanical

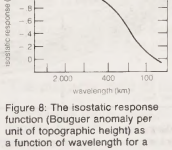


Figure 8: The isostatic response function (Bouguer anomaly per unit of topographic height) as a function of wavelength for a typical continental area.

rather than chemical properties, may have only a very small discontinuity in density associated with it.

In the light of current ideas on plate tectonics, a fundamental question is whether, with a rigid lithosphere, isostatic compensation exists on the local point-by-point basis visualized by Pratt and Airy. Should not the application of loads to the lithosphere cause it to flex gently as an elastic plate floating on a dense fluid? In the extreme case, even though most of the signal in the gravity anomaly is produced at the Mohorovičić discontinuity, that interface is passively warped into surfaces that parallel, and are determined by, the lithosphere as a whole. The isostatic response function $Q(k)$ can be expressed for this model also. In contrast to equation (25), it contains two model parameters: depth of compensation and flexural rigidity of the lithosphere. Various continental areas have been investigated using this model, with the result that very small values have been found for the thickness of the elastic lithosphere, often less than the depth to the Mohorovičić discontinuity, which is typically 35 kilometres beneath continents. For example, in the case of eastern North America, the method shows an elastic thickness of only 20 kilometres, much less than the lithospheric thickness of 50 to 100 kilometres suggested by seismic wave velocities. While it is conceivable that the effective thickness of "rigid" material is different for different time scales, the result contains a paradox. For if the roots are in the asthenosphere below the base of the lithosphere, there is no reason why they should not have disappeared by flow in the case of old topographic features. The solution seems to lie in the fact that the Mohorovičić discontinuity does not simply flex in a passive way when mountains and other relief features are formed. Horizontal forces fracture and thicken the crust by overthrusting, as shown in seismological profiles of the Himalayas. Compensation at the base of the crust is produced independently of lithospheric flexure, which itself is determined by the uncompensated, rather than total, topographic load.

The global significance of gravity anomalies. The investigations of isostasy described above are statistical in approach in that they are directed toward determining depth of compensation and lithospheric thickness from large data sets without reference to individual geologic structures. On the other hand, the study of gravity is capable of extending the understanding of such major features of global plate tectonics as mid-oceanic ridges and subduction zones. Just as plate tectonics itself began with a consideration of features of the ocean floors, it is the gravitational field over the oceans that is most diagnostic for these applications.

Subduction zones. It has been pointed out that, in comparison to magnetic and seismological studies, gravity measurements played only a secondary role in establishing the plate-tectonic theory in the 1960s. On the other hand, as early as 1929 the Dutch geophysicist Felix A. Vening Meinesz had shown that some of the most spectacular negative gravity anomalies on Earth are associated with oceanic trenches (Figure 9). These regions are now known to be subduction zones—*i.e.,* plate margins along which a slab of oceanic lithosphere is forced to descend beneath a continental plate. The narrow negative gravity anomalies result from the accumulation of light sedimentary material trapped between the plates. Not obvious on the map but more significant from the point of view of global tectonics is the extensive positive anomaly landward from the trough. These can be interpreted on the basis of a small positive density contrast between the descending slab and the surrounding normal mantle. Since the slab is similar chemically to its surroundings, the density contrast results from the lower temperature of the slab. Detailed analyses of the gravity field over subduction zones, in conjunction with seismic measurements, thus yield information on slab geometry, depth extent, and temperature.

Negative gravity anomalies

Mid-oceanic ridges. These structures represent the spreading axes of plate tectonics, along which hot volcanic material is ejected to become new oceanic lithosphere. Bathymetrically, the ridges are striking topographic features, rising one to three kilometres above the deep ocean floor on either side. That they are associated with geoidal highs can be seen on the global map of the geoid (Figure 4). On the average, however, the geoidal effect of such a ridge is moderate (Figure 10), indicating that some form of compensation is operative. Temperature variations in the oceanic plates moving away from the ridge axis appear to play a major role in the compensation. The ejected material is hot and therefore of low density. As the plates move apart, they lose heat vertically to the ocean above, cool, and become denser. If it is assumed that isostatic compensation is complete (it is a Pratt-type of isostasy), the profile of the ocean floor can be computed and is found to be in good agreement with the bathymetry. This lends strength to the cooling-plate model of ocean ridge topography. Nevertheless, if isostatic compensation is complete, why should there be a gravity high, or equivalently, a geoidal

Geoidal highs

Adapted from W.A. Heiskanen and F.A. Vening Meinesz, *The Earth and Its Gravity Field*; McGraw-Hill Co., 1958

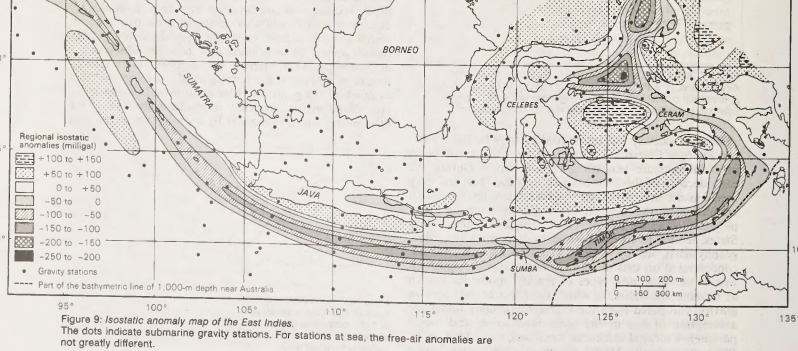


Figure 9: *Isostatic anomaly map of the East Indies.*
The dots indicate submarine gravity stations. For stations at sea, the free-air anomalies are not greatly different.

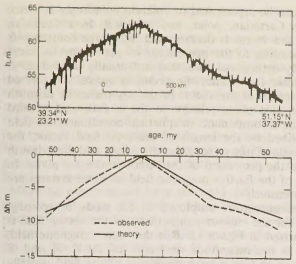


Figure 10: (Top) Observed geoidal profile across the Mid-Atlantic Ridge. (Bottom) The smoothed observed profile as compared with that calculated for spreading plates that are thermally compensated.
After W.F. Haxby and D.L. Turcotte, *Journal of Geophysical Research*, no. 83, p. 5475 (1978); copyright American Geophysical Union

high? The answer lies in the vertical separation between the positive masses (*i.e.*, the ridge itself relative to deep ocean) and the compensating "negative masses" that are distributed through the hot plates and therefore at greater depth. The result is similar to a "double layer" of opposite poles in magnetism. Positive gravitational potential is produced at the ocean surface, leading to the outward geoidal bulge. As Figure 10 indicates, both observations and model calculations indicate a geoidal height that decreases almost linearly with the age of the seafloor. The result is important for the interpretation of the oceanic geoid because it allows the effect of ridges to be removed so as to better display the effect of deeper structures.

There is another type of superficial feature that produces the small isolated circular highs readily visible on the geoid over the Pacific Ocean—namely, the seamount, or submarine volcanic cone, built on the ocean floor. As in the case of ridges, these features, though appearing to be compensated by plate flexure, produce geoidal highs because of the separation in depth of their masses and the compensation. Most interesting is the difference in signature between seamounts on young and older oceanic lithosphere. In the former case, the lithospheric plate has low flexural rigidity, and so it bends in such a way as it achieves more local compensation; the geoidal high is of small amplitude and is flanked by lows. On cold, stronger lithosphere, the compensation is of such regional extent that a pronounced geoidal high appears.

Mantle convection. Convection of mantle material, at velocities of centimetres per year, has been suggested as the driving force for plate motions, and it is natural to look for some expression of a system of mantle convection currents in the gravitational field. In a thermally driven convecting system, heated mantle material of lower density rises, moves horizontally dragging the lithosphere with it, and, after cooling, descends to complete the cycle. At first sight, one might expect that areas of rising currents would produce gravity lows because of the density deficiency, but in a system with a free upper boundary the convection currents themselves deform the boundary vertically, "piling up" material above rising currents. Model calculations show that in most cases this boundary effect predominates over the density variations in the material, leading to positive gravity or geoidal anomalies over rising currents and to negative ones over sinking currents.

While there is still no proof that mantle convection exists, there are gravity anomalies of intermediate-length scale that are suggestive of a system of convection cells. Figure 11 shows the geoid for the Pacific Ocean with short-wavelength features smoothed out. The effect of the East Pacific Rise has not been removed and appears as a high toward the lower right of the figure. Most striking, however, are the northwesterly alignments of alternating geoidal highs and lows (*e.g.*, D, T, G, V) extending in the direction of plate motion. The fact that these features of the geoid correlate positively with the bathymetry supports the suggestion that both result from some system of convection cells in the upper mantle.

Possibility of convection cells in the mantle

Areas of glacial depression and postglacial rebound. Prominent negative areas of the geoid (Figure 4) include Antarctica and northern North America centred over Hudson Bay. Both of these regions were covered with ice sheets during Pleistocene time (about 10,000 to 1,700,000 years ago), and it is believed the negative anomalies are related to a residual depression of the lithosphere that remained after the ice retreat. When an ice sheet develops on a continental area, compensation is provided by the downward deflection of the continental lithosphere, leading to outflow of the asthenospheric material beneath it. Melting of the ice leaves a depressed crust, which rebounds gradually as the viscous material flows back. During the period of rebound, the area will display a geoidal low and a negative free-air anomaly of a magnitude (over a broad region) of $2\pi G\rho y$, where ρ is the actual density of asthenospheric material (about 3.3 grams per cubic centimetre) and y is the distance that the Earth's crust must still rise. Gravity measurements thus yield values of y (it is roughly 200 metres for Hudson Bay). Taken in conjunction with the present-day rate of uplift, as given by repeated precise leveling, they allow the viscosity of the mantle material to be estimated.

Because these negative regions are near the poles, they contribute strongly to the term in J_2 in the potential (see above) and therefore to the flattening of the ellipsoid. Evidence exists from the perturbation of some satellite orbits—notably the orbit of the geodetic satellite Lageos—that J_2 is decreasing with time. There is a motion that cannot be explained by tidal or other known effects but that is consistent with a rate of decrease of J_2 of $\dot{J}_2 = -(26 \pm 6) \times 10^{-12}\ \text{yr}^{-1}$. The fact that the flattening and therefore J_2 are decreasing with time is consistent with the upward motion of the previously glaciated polar regions. (G.D.G.)

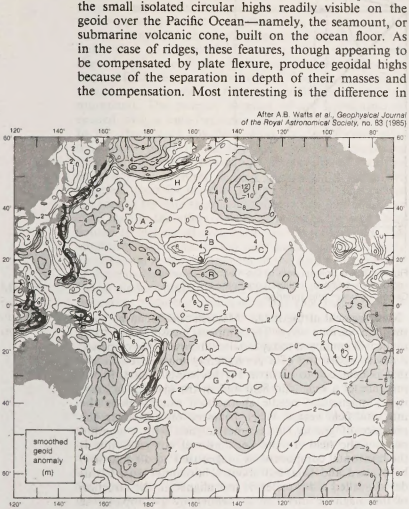


Figure 11: A smoothed geoid map of the Pacific Ocean. The intervals between contours are two metres. The shaded areas represent negative geoidal anomalies, and the letters indicate individual geoidal highs and lows.

The magnetic field of the Earth

The Earth's steady magnetic field is produced by many sources, both above and below the planet's surface. From the core outward, these include the geomagnetic dynamo, crustal magnetization, ionospheric dynamo, ring current,

magnetopause current, tail current, field-aligned currents, and auroral electrojets. The geomagnetic dynamo is the most important source because, without the field it creates, the other sources would not exist. Not far above the Earth's surface the effect of other sources becomes as strong or stronger than that of the geomagnetic dynamo. In the discussion that follows, each of these sources is considered and their respective causes explained.

Major disturbances of the main magnetic field

The Earth's magnetic field is subject to variation on all time scales. Each of the major sources of the so-called steady field undergo changes that produce transient variations, or disturbances. The main field has two major disturbances: quasi-periodic reversals and secular variation. The ionospheric dynamo is perturbed by seasonal and solar cycle changes as well as by solar and lunar tidal effects. The ring current responds to the solar wind (the ionized atmosphere of the Sun that expands outward into space and carries with it the solar magnetic field), growing in strength when appropriate solar wind conditions exist. Associated with the growth of the ring current is a second phenomenon, the magnetospheric substorm, which is most clearly seen in the aurora borealis. An entirely different type of magnetic variation is caused by magnetohydrodynamic (MHD) waves. These waves are sinusoidal variations in the electric and magnetic fields that are coupled to changes in particle density. They are the means by which information about changes in electric currents is transmitted, both within the Earth's core and in its surrounding environment of charged particles (see below).

OBSERVATIONS OF THE EARTH'S MAGNETIC FIELD

Representation of the field. Electric and magnetic fields are produced by a fundamental property of matter, electric charge. Electric fields are created by charges at rest relative to an observer, whereas magnetic fields are produced by moving charges. The two fields are different aspects of the electromagnetic field, which is the force that causes electric charges to interact. The electric field, **E**, at any point around a distribution of charge is defined as the force per unit charge when a positive test charge is placed at that point. For point charges the electric field points radially away from a positive charge and toward a negative charge.

A magnetic field is generated by moving charges—*i.e.*, an electric current. The magnetic induction, **B**, can be defined in a manner similar to **E** as proportional to the force per unit pole strength when a test magnetic pole is brought close to a source of magnetization. It is more common, however, to define it by the Lorentz-force equation. This equation states that the force felt by a charge q, moving with velocity $\boldsymbol{v}$, is given by

Lorentz-force equation

$$F = q(\boldsymbol{v} \times \boldsymbol{B}). \tag{26}$$

In this equation bold characters indicate vectors (quantities that have both magnitude and direction) and nonbold characters denote scalar quantities such as B, the length of the vector $\boldsymbol{B}$. The $\times$ indicates a cross product (*i.e.*, a vector at right angles to both $\boldsymbol{v}$ and $\boldsymbol{B}$, with length $vB \sin \theta$). Theta is the angle between the vectors $\boldsymbol{v}$ and $\boldsymbol{B}$. (B is usually called the magnetic field in spite of the fact that this name is reserved for the quantity H, which is also used in studies of magnetic fields.) For a simple line current, the field is cylindrical around the current. The sense of the field depends on the direction of the current, which is defined as the direction of motion of positive charges. The right-hand rule defines the direction of $\boldsymbol{B}$ by stating that it points in the direction of the fingers of the right hand when the thumb points in the direction of the current.

In the International System of Units (SI), the electric field is measured in terms of the rate of change of potential, volts per metre (V/m). Magnetic fields are measured in units of tesla (T). The tesla is a large unit for geophysical observations and a smaller unit, the nanotesla (nT: one nanotesla equals 10^{-9} tesla), is normally used. A nanotesla is equivalent to one gamma, a unit originally defined as 10^{-5} gauss, which is the unit of magnetic field in the centimetre-gram-second system. Both the gauss and gamma are still frequently used in the literature on geomagnetism even though they are no longer standard units.

Both electric and magnetic fields are described by vectors, which can be represented in different coordinate systems, such as Cartesian, polar, and spherical. In a Cartesian system the vector is decomposed into three components corresponding to the projections of the vector on three mutually orthogonal axes that are usually labeled x, y, z. In polar coordinates the vector is typically described by the length of the vector in the x-y plane, its azimuth angle in this plane relative to the x axis, and a third Cartesian z component. In spherical coordinates the field is described by the length of the total field vector, the polar angle of this vector from the z axis, and the azimuth angle of the projection of the vector in the x-y plane. In studies of the Earth's magnetic field all three systems are used extensively.

The nomenclature employed in the study of geomagnetism for the various components of the vector field is summarized in Figure 12. $\boldsymbol{B}$ is the vector magnetic field, and F is the magnitude or length of $\boldsymbol{B}$. X, Y, and Z are the three Cartesian components of the field, usually measured with respect to a geographic coordinate system. X is northward, Y is eastward, and completing a right-handed system, Z is vertically down toward the centre of the Earth. The magnitude of the field projected in the horizontal plane is called H. This projection makes an angle D (for declination) measured positive from the north to the east. The dip angle, I (for inclination), is the angle that the total field vector makes with respect to the horizontal plane and is positive for vectors below the plane. It is the complement of the usual polar angle of spherical coordinates. (Geographic and magnetic north coincide along the "agonic line.")

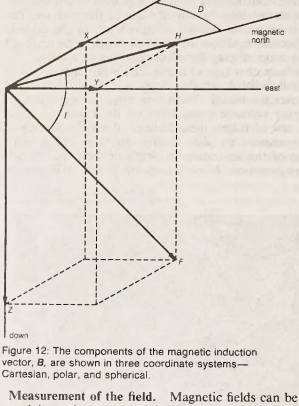


Figure 12: The components of the magnetic induction vector, **B**, are shown in three coordinate systems—Cartesian, polar, and spherical.

Measurement of the field. Magnetic fields can be measured in various ways. The simplest measurement technique still employed today involves the use of the compass, a device consisting of a permanently magnetized needle that is balanced to pivot in the horizontal plane. In the presence of a magnetic field and in the absence of gravity, a magnetized needle aligns itself exactly along the magnetic field vector. When balanced on a pivot in the presence of gravity, it becomes aligned with a component of the field. In the conventional compass, this is the horizontal component. A magnetized needle may also be pivoted and balanced about a horizontal axis. If this device, called a dip meter, is first aligned in the direction of the magnetic meridian as defined by a compass, the needle lines up with the total field vector and measures the inclination angle I. Finally, it is possible to measure the magnitude of the horizontal field by the oscillations of the compass needle. It can be shown that the period

of such an oscillation depends on properties of the needle and the strength of the field.

Magnetic observatories continuously measure and record the Earth's magnetic field at a number of locations. In an observatory of this sort, magnetized needles with reflecting mirrors are suspended by quartz fibres. Light beams reflected from the mirrors are imaged on a photographic negative mounted on a rotating drum. Variations in the field cause corresponding deflections on the negative. Typical scale factors for such observatories correspond to 2–10 nanoteslas per millimetre vertically and 20 millimetres per hour horizontally. A print of the developed negative is called a magnetogram.

Magnetic observatories have recorded data in this manner for well over 100 years. Their magnetograms are photographed on microfilm and submitted to world data centres, where they are available for scientific or practical use. Such applications include the creation of world magnetic maps for navigation and surveying; correction of data obtained in air, land, and sea surveys for mineral and oil deposits; and scientific studies of the interaction of the Sun with the Earth.

Measurements with proton-precession and flux-gate magnetometers

In recent years, other methods of measuring magnetic fields have proved more convenient, and older instruments are being gradually replaced. One such method involves the proton-precession magnetometer, which makes use of the magnetic and gyroscopic properties of protons in a fluid such as gasoline. In this method, the magnetic moments of protons are first aligned by a strong magnetic field produced by an external coil. The magnetic field is then turned off abruptly, and the protons try to align themselves with the Earth's field. However, since the protons are spinning as well as magnetized, they precess around the Earth's field with a frequency dependent on the magnitude of the latter. The external coil senses a weak voltage induced by this gyration. The period of gyration is determined electronically with sufficient accuracy to yield a sensitivity between 0.1 and 1.0 nanotesla.

An instrument that complements the proton-precession magnetometer is the flux-gate magnetometer. In contrast to the proton-precession magnetometer, the flux-gate device measures the three components of the field vector rather than its magnitude. It employs three sensors, each aligned with one of the three components of the field vector. Each sensor is constructed from a transformer wound around a core of high-permeability material (*e.g.*, mu-metal). The primary winding of the transformer is excited with a high-frequency (~5 kilohertz) sine wave. In the absence of any field along the transformer axis, the output signal in the secondary winding consists of only odd harmonics (component frequencies) of the drive frequency. If, however, a field is present, it biases the hysteresis loop for the core in one direction. This causes the core to become saturated sooner in one half of a drive cycle than in the other. This in turn causes the secondary voltage to include all even harmonics as well as odd. The amplitude and phase of the even harmonics are linearly proportional to the component of the field along the axis of the transformer.

Modern magnetic observatories

Most modern magnetic observatories have both a proton-precession magnetometer and a flux-gate magnetometer mounted on granite pillars in nonmagnetic, temperature-controlled rooms. The outputs from the instruments are electrical signals, and they are digitized and recorded on magnetic media. Many observatories also transmit their data soon after acquisition to central facilities where they are stored with data from other locations in a large computer database.

Magnetic measurements are often made at locations remote from fixed observatories. Such measurements are commonly part of a survey designed to better define the Earth's main field or to detect anomalies in it. Surveys of this type are routinely carried out by foot, ship, aircraft, and spacecraft. For surveys near the Earth's surface the proton-precession magnetometer is almost always used because it does not need to be precisely aligned. Above the Earth's surface the main field decreases rapidly, and the need for precise alignment is less severe. Thus, flux-gate magnetometers are generally employed on spacecraft. Calculation of components of the vector field in a coordinate system fixed with respect to the Earth requires knowledge of the location and orientation of the spacecraft.

CHARACTERISTICS OF THE EARTH'S MAGNETIC FIELD

To a first approximation the magnetic field observed at the surface of the Earth is like that of a magnet aligned with the planet's rotation axis. Figure 13 shows such a field for a bar magnet located at the centre of a sphere. If the sphere is taken to be the Earth with the north geographic pole at the top of the diagram, the magnet must be oriented with its north magnetic pole downward toward the south geographic pole. Then, as shown in the diagram, magnetic field lines leave the north pole of the magnet and curve around until they cross the Earth's Equator pointing geographically northward. They curve still more reentering the Earth in northern latitudes, finally returning to the south pole of the magnet. At the present time, the north geographic pole corresponds to the south pole of the equivalent bar magnet. This has not always been the case. Many times in the history of the Earth the direction of the equivalent magnet has pointed in the opposite direction (see below *Reversals in the main field*).

Magnetic field lines

The magnetic field lines shown in the diagram are not real entities although they are frequently treated as such. A magnetic field is a continuous function that exists at every point in space. A field line is simply a means for visualizing the direction of this field. It is defined as a curve in three dimensions that is everywhere tangential to the local magnetic field. The pattern of field lines created by a bar magnet is called a dipolar field because it has the same shape as the electric field produced by two (di-) slightly separated charges (poles) of opposite sign. The dipole field of the Earth is, of course, not produced by a bar magnet at its centre. As will be discussed later, it is instead produced by electric currents within the Earth's liquid core. To produce the present field, the equivalent current must be a westward equatorial loop as shown in Figure 13. In SI units the dipole moment, μ, for the Earth is 7.95×10^{22} A/m^2 (amperes per square metre). Since $\mu = IA$ (current times area), a loop the size of the liquid core ($R_c = 3.48 \times 10^6$ m) would require an equivalent current of nearly 2×10^9 A.

It can be seen from Figure 13 that the magnetic field of a dipole is vertical along the polar axis and horizontal along the equator. These properties lead to definitions of equator and pole in the Earth's more complex field. Thus the geomagnetic equator is defined as the line around the Earth's surface where the actual field is horizontal. Similarly, the magnetic dip poles are the two points at which the field is vertical. If observations are extended above or below the surface, the location of the equator is a surface (planar for a dipole) and the poles lie along curves.

At a given distance in a pure dipole field, the polar field is always twice the equatorial field. The map in Figure 14 demonstrates that this is roughly true for the Earth's field. The map shows contours of constant total field magnitude according to a 1980 model plotted on a geographic

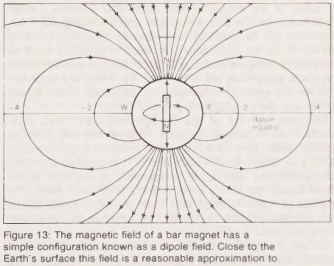


Figure 13: The magnetic field of a bar magnet has a simple configuration known as a dipole field. Close to the Earth's surface this field is a reasonable approximation to the actual field.

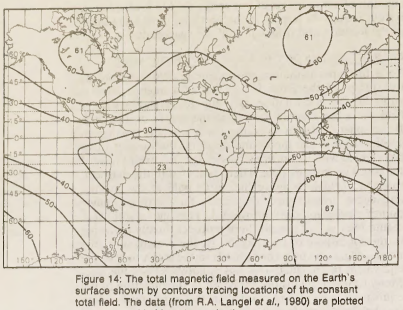


Figure 14: The total magnetic field measured on the Earth's surface shown by contours tracing locations of the constant total field. The data (from R.A. Langel *et al.*, 1980) are plotted on a geographic Mercator projection.

Adapted from W.D. Parkinson, *Introduction to Geomagnetism*, Scottish Academic Press, 1983; original data published by the American Geophysical Union, 1980.

Mercator projection. The largest fields occur at two points in the Northern and Southern hemispheres not far from the geomagnetic poles. The weakest field occurs along the magnetic equator, with the lowest value being observed on the Atlantic coast of South America.

Several facts about the Earth's field are apparent from the total field map. First, the dipole approximating it is not exactly aligned with the rotation axis. The poles of the dipole are located roughly in northern Canada and on the coast of Antarctica rather than at the geographic poles. This implies that the dipole is tilted away from the rotation axis in a geographic meridian passing through the eastern United States. The exact tilt of the best centred dipole is 11° away from the geographic North Pole toward North America at a longitude 71° W of Greenwich. The total field map also suggests that the field is not exactly centred in the Earth, for if it were the field strength should be nearly constant along the Equator.

Relationship of the dipole to the geomagnetic poles

The mathematical description of a vector field on the surface of a sphere is quite complicated. In studies of the Earth's field it is usually done by multipole expansions. The field is assumed to be made of the superposition of fields from a series of poles located at the centre of the Earth. The first pole in this expansion is a monopole corresponding to only one pole of a magnet. Since no magnetic monopole has ever been observed, this term is not used. The next term is the dipole, then the quadrupole, and so forth. When the Earth's field is described in this manner, it is found that the dipole term accounts for more than 90 percent of the field. If the contribution from a centred dipole is subtracted from the observed field, the residual is called the non-dipole field, or regional geomagnetic anomaly.

Current maps of the regional anomaly for various components of the magnetic field show that there is a large maximum in the South Atlantic and in Mongolia. This anomaly can be partially explained by offsetting the best fit dipole in an appropriate manner. Anomalies such as this affect compass readings in polar regions and influence particles trapped in the outer field. They also are responsible for the separation between the locations of the dipole poles and the geomagnetic poles.

Magnetic surveys of the Earth's field have been conducted with increasing accuracy for well over 100 years. In recent times, they have been conducted on approximately a 10-year schedule. For each survey it is possible to define the dipole and non-dipole components of the field. It has been found that both change systematically with time. The nature of these changes and their probable explanations are discussed below in *Sources of variation in the steady magnetic field.*

In the multipole description of the Earth's field, it is shown that the effects of higher order poles decrease more rapidly with distance than those of the lower order poles. The field of a monopole, for example, decreases as the inverse square of distance, the dipole as the inverse cube, and so on. Because of this property, it might be expected that the outer portions of the Earth's field would be almost purely dipolar. Recent spacecraft observations, however, show that this is not true. The field departs radically from that of a dipole at altitudes of only a few Earth radii.

Surface observations do not suggest that significant distortion of the Earth's field should occur close to the planet. The technique of multipole expansion makes it possible to separate the observed surface field into parts of origin internal and external to the Earth. When surface observations are averaged over several years, less than 1 percent of the surface field is produced by external sources. Thus the existence of the external distortion is surprising.

Spacecraft determination of the Earth's outer field

The actual configuration of the Earth's outer magnetic field as recently determined by spacecraft is summarized in Figure 15. The diagram shows projections of magnetic field lines into the noon–midnight meridian at a time near an equinox. At this time the Earth's rotation axis is perpendicular to the Earth–Sun line. The dipole axis will be tilted another plus or minus 11°, depending on the time of day. On the dayside of the Earth, the magnetic field of the planet terminates at a distance of about 10 R_e (where R_e is the Earth's equatorial radius of about 6,378 kilometres). The boundary that exists at this point is called the magnetopause (break in magnetic field). Outside this boundary magnetic fields and particles are present, but they belong to the Sun's atmosphere and not to the Earth's. On the nightside, the magnetic field is drawn out into a long tail consisting of two lobes separated by a 14-R_e-thick sheet of particles called the plasma sheet. The plasma sheet has an inner boundary about 11 R_e behind the Earth. It also has upper and lower boundaries as shown. The projection of these boundaries onto the northern and southern portions of the atmosphere at about 67° magnetic latitude corresponds to two regions called the nightside auroral ovals. The aurora borealis and aurora australis (northern lights and southern lights) appear within the regions defined by the feet of these field lines and are caused by bombardment of the atmosphere by energetic charged particles. On the dayside, magnetic field lines from high latitudes split, some crossing the Equator, while others cross over the polar caps. The regions where the field lines split are called polar cusps. The projection of the polar cusps on the atmosphere at about 72° magnetic latitude creates the dayside auroral ovals. Auroras can be seen in these regions in the dark hours of winter, but they are much weaker than on the nightside because the particles that produce them have much less energy. The projections of the two lobes of the magnetic tail onto the atmosphere are the polar caps.

Within the middle of the Earth's field are several other important boundaries and regions that cannot be detected by magnetic field observations. Close to the Earth (1–2 R_e) is the inner Van Allen radiation belt, which consists

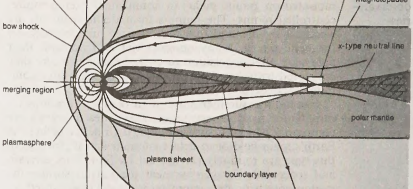


Figure 15: Field lines of the Earth's magnetic field are shown projected into the noon–midnight meridian plane. Major features of the Earth's magnetic field are labeled.

of very energetic particles created by cosmic rays. Centred at about 4–5 R_e is the outer Van Allen belt, created from charged particles of both solar and atmospheric origin. Also at this distance is the plasmapause. This is a boundary in the Earth's plasma (a relatively cold gas consisting of equal numbers of electrons and positive ions) and, as such, actually constitutes a boundary in the planet's electric field.

SOURCES OF THE STEADY MAGNETIC FIELD

The geomagnetic dynamo. Observations of the magnetic field of the Earth's surface indicate that more than 90 percent of this field arises from sources internal to the planet. A variety of mechanisms for generating this field have been proposed, but at present only the geomagnetic dynamo is seriously considered. In the dynamo mechanism, fluid motion in the core moves conducting material across an existing magnetic field and creates an electric current. This current produces a magnetic field that also interacts with the fluid motion to create a secondary magnetic field with the same orientation as the original field. The two fields together are stronger than the original. The additional energy in the amplified field comes at the expense of a decrease in energy in the fluid motion.

Dynamo mechanism

Thermal heating in the core is the process that drives fluid motion. For many years it was thought that this heating was caused by radioactive elements dissolved in the liquid core. Recent work suggests that freezing of the liquid core is more important. Seismic studies have shown that the centre of the Earth is a solid sphere of iron with an approximate radius of 1,200 kilometres. This sphere is surrounded by an outer core of liquid iron. With time, the inner surface of the liquid core freezes onto the outer surface of the solid core. Energy released in the freezing process heats the surroundings to a high temperature. The heat flows in all directions, raising the temperature of adjacent regions. Because heat cannot be lost from the interior, it eventually flows to the surface. There, it is radiated into the cold of space as infrared radiation. This process establishes a radial temperature distribution that decreases toward the surface. If heat is generated too rapidly for conduction to carry it away, a second process, convection, becomes important. In convection, energy is transported by bubbles of hot fluid that rise toward cooler regions carrying more heat than flows through the same material at rest.

Several conditions must be satisfied for the fluid motion to produce a magnetic field. First, the fluid must be electrically conducting. Second, a magnetic field must be present, possibly as a relict of the initial formation of the body. Third, some force must introduce twists into the fluid motion so that the initial magnetic field becomes distorted by the motion. For the Earth, liquid iron is conducting, an initial magnetic field is likely, and the Coriolis force introduces twists. The Coriolis force is the force felt by a fluid in or on a rotating body. It is the force that creates cyclonic storms in the Earth's atmosphere, and in the Northern Hemisphere it causes a fluid rising radially to rotate counterclockwise.

Generation of a self-sustaining magnetic field

The example presented in Figure 16, designated the $a\omega$ dynamo, illustrates how these factors might generate a self-sustaining magnetic field. Assume first (A) that there is present an initial poloidal magnetic field (one lying in meridian planes). Suppose next that the innermost part of the field line is embedded in a fluid rotating more rapidly than the outer parts of the fluid. In good conductors, magnetic field lines are nearly frozen into the fluid and have to move as the fluid moves. After many rotations a field line will "wrap up" around the rotation axis, creating a large toroidal field (one lying in planes perpendicular to the rotation axis). Since the conductivity is not perfect, the toroidal loop may diffuse through the fluid, disconnecting itself from the original poloidal field (B). This process is called the omega effect because it depends on the rotational velocity of the fluid.

Next, consider the effect of radial fluid motion on the toroidal field. At various points in the liquid core, fluid is rising in cells driven by thermal convection. The rising fluid carries with it the toroidal magnetic field. As it rises, the Coriolis force deflects the fluid and causes it to spin around the central axis of the cell, thereby twisting the magnetic field. After a rotation of about 270° the magnetic field lines begin to twist about themselves and can diffuse through the conductor, disconnecting from the toroidal loop (C). At this stage, the rising loop is oriented in a meridian plane with the field pointing in the same direction as the original field—*i.e.*, poloidal. This process is called the alpha effect (because the effects are proportional with constant, α, to the background field). Finally, small loops may merge into a single large loop, recreating the initial poloidal field (D). In cells of sinking fluid, the toroidal field wraps in the opposite direction and the poloidal loops have the opposite polarity. If the sinking process were exactly symmetrical, field loops produced in this manner would cancel loops created by rising fluid. Thus, for the process to create a net field of the correct sign, loops produced by sinking fluid must be weaker than loops resulting from rising fluid.

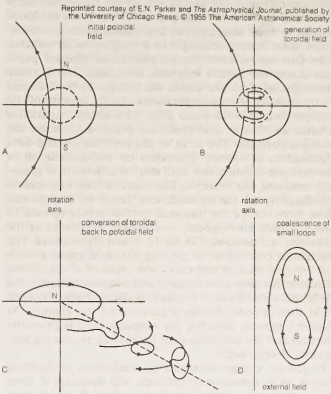


Figure 16: Generation of a self-sustaining magnetic field might be accomplished by this sequence of processes called the $a\omega$ dynamo.

As discussed above, the simplest possible poloidal magnetic field is dipolar. Such a field could be produced by a single loop of electric current circulating around the Earth's rotation axis in the equatorial plane. The slight electric resistance of the conducting Earth, however, would long ago have dissipated this current if it were not continuously regenerated. As the illustration makes clear, this generation process is complex and depends on both radial motion and rotation of the fluid core.

Crustal magnetization. Magnetic fields measured at the Earth's surface are not entirely produced by the internal dynamo. Radially outward from the Earth's core, the next major source of magnetic field is crustal magnetization. The temperature of the materials constituting the crust is cool enough for them to exist in solid form. The solids may become magnetized by the Earth's main field and cause detectable anomalies.

Induced and remanant magnetization

Crustal magnetization is of two types: induced and remanant. Induced magnetization occurs when the elementary magnetic dipoles of crustal materials are aligned by the Earth's main field, just as a compass needle is aligned. If a material of particularly high susceptibility to magnetization is concentrated, as in a mineral deposit, it also can be approximated as a bar magnet that creates a small dipole field. On the scale of such concentrations, the Earth's main field is uniform, so, depending on an observer's location relative to the small dipole, its field may either add to or subtract from the main field. Because induced magnetiza-

tion is proportional to the strength of the inducing field, it vanishes when the primary field vanishes.

Remanant magnetization is similar to induced magnetization in that it is produced in a material by a primary field, but once created it persists after the primary field has disappeared. The phenomenon depends on the presence of ferromagnetic materials that form "magnetic domains," regions of aligned dipoles held in place by interatomic forces. In the Earth's crust, most remanant magnetization is created by trapping the dipole alignment of the Earth's main field as molten rocks harden.

The ionospheric dynamo. Above the Earth's surface, the next source of magnetic field is the ionospheric dynamo—an electric current system flowing in the planet's ionosphere. Beginning at about 50 kilometres and extending above 1,000 kilometres with a maximum at 400 kilometres, the ionosphere is formed primarily by the action of sunlight on atmospheric particles. There, sunlight strips electrons from neutral atoms and produces a partially ionized gas (plasma). On the dayside of the Earth near local noon and near the subsolar point, the Sun heats the ionosphere to high temperatures and causes it to flow away from noon toward midnight in a roughly radial pattern. The flow moves both neutral atoms and charged particles across the Earth's magnetic field lines. The Lorentz force discussed earlier causes the charges to be deflected in opposite directions perpendicular to the velocity of the charges and also the local field. This charge separation creates an electric field that also exerts a force on the charged particles. The form of the resulting electric field distribution is strongly dependent on the distribution of ionospheric conductivity and magnetic field. It is generally assumed, for example, that there is little ionospheric conductivity on the nightside and hence no current can flow there. As for the magnetic field, it points upward in the Southern Hemisphere, horizontally northward at the Equator, and downward in the Northern Hemisphere. The horizontal component of the magnetic field exerts a vertical force on charges that move as a result of winds. At the Equator this causes the positive and negative charges to be deflected vertically and produces a strong vertical electric field that impedes further separation of the charges. At higher magnetic latitudes, the magnetic field is primarily vertical and the deflections are horizontal, producing horizontal electric fields.

In general, charges separated by mechanical or chemical forces, as in dynamos or batteries, will discharge if there is an external electrical conductor through which they can flow. At high and low latitudes this process occurs in the same medium that generates the charge separation. The actual current path is particularly complex in the ionosphere because the electrical conductivity is spatially inhomogeneous and anisotropic—*i.e.*, it varies from point to point and has different values in different directions relative to the magnetic and electric fields present.

Complexity of the ionospheric dynamo current system

The form of the electric currents flowing in the ionosphere has been deduced from ground observations of daily variations in the magnetic field. On magnetically quiet days the field is observed to change in a systematic manner dependent primarily on local time and latitude. This variation has been dubbed the solar quiet-day variation, S_q. The magnetic variations can be used to deduce an equivalent electric current system, which, if flowing in the E region of the ionosphere, would produce the observed changes. This system is shown in Figure 17 for the equinoctial conditions of equal illumination of both hemispheres when the pattern is symmetrical about the Equator. The pattern consists of two current vortices circulating about foci at $+$ and $-30°$ magnetic latitude. Viewed from the Sun, circulation is counterclockwise in the Northern Hemisphere and clockwise in the Southern Hemisphere. Approximately 500,000 amperes flow eastward parallel to the Equator between the two foci. Apart from small changes brought about by daily rotation of small anomalies in the main field, the current and its effects at a fixed point in space are nearly steady. A magnetic observatory, however, rotates beneath different parts of the current system and records a time-varying magnetic field.

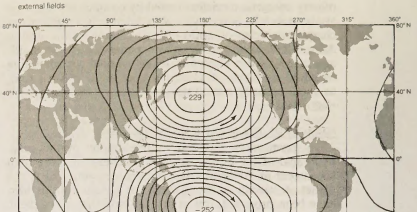


Figure 17: The ionospheric dynamo current system as seen from the Sun at an equinox. The contour lines are at intervals of 24.5 kiloamperes.

From W.D. Parkinson, *Introduction to Geomagnetism*, Scottish Academic Press Ltd. (1983), reproduced with permission of the director, Bureau of Mineral Resources, Geology and Geophysics, Canberra

A detailed analysis of the daily variation reveals that several important factors contribute to the ionospheric wind system driving the dynamo. The most significant of these is the solar heating of the atmosphere discussed above. There is, however, a semidiurnal component caused by solar gravity that is roughly half as large as the diurnal component. As in the oceans, the tidal effect of gravity produces peaks in pressure at midnight as well as at noon. The resulting winds are more complex than is the case for the diurnal component. Similarly, there is a semidiurnal lunar component driven by lunar gravity. This variation is named the lunar daily variation, L. Its peak-to-peak amplitude is about $1/20$ that of S_q.

The ring current. Farther out, at 4 R_e and beyond, is the next major source of magnetic field, the ring current. At this distance almost all atmospheric particles are fully ionized and, hence, subject to the effects of electric and magnetic fields. Furthermore, the density of the particles is so low that the time between collisions may be many days or months. Here, energetic charged particles tend to behave independently rather than as part of a fluid. The behaviour of these particles may be approximated by the superposition of three types of motion, as shown schematically in Figure 18. These types include gyration about the

Particle motions

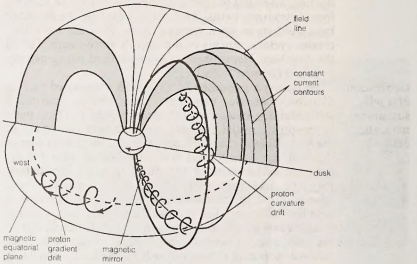


Figure 18: The motion of single particles in the Earth's magnetic field may be approximated by the superposition of their gyration about the main field, "bounce" along the field lines, and azimuthal drift in rings around the Earth. The trajectories of individual particles in the ring current fill a doughnut-shaped volume of space. The current produced by the particle drift causes a decrease in the surface field (see text).

main field, "bounce" along field lines, and azimuthal drift in rings around the Earth.

Gyration is caused by the Lorentz force, which makes charged particles move in circles around magnetic field lines. Reflection of particles at the ends of field lines is produced by the converging geometry of a dipole field. As a gyrating charged particle approaches the Earth moving along a field line, the particle encounters a magnetic mirror that reflects it. The mirror force is a component of the Lorentz force antiparallel to the motion of the particle when field lines converge.

Azimuthal drift is produced by two effects: a decrease in the strength of the main field away from the Earth, and a curvature of magnetic field lines. The first effect is easy to understand by considering the dependence of the particles' radius of gyration on the strength of the magnetic field. Strong fields cause small orbits. When a particle gyrates in the Earth's field, it has a larger radius close to the Earth than it does farther away. The projection of such motion into the equatorial plane is a cycloidal trajectory in a ring around the Earth rather than a simple circle around a local field line. Particles of opposite charge drift in opposite directions because their sense of gyration about the direction of the magnetic field is opposite—*i.e.*, protons gyrate in a left-handed sense (left handed with respect to the Earth's rotation axis) and drift westward, while electrons gyrate in a right-handed sense and drift eastward. Because the particles drift in opposite directions, they produce an electric current in the same direction as the proton drift.

A second cause of azimuthal drift known as curvature drift is also depicted in Figure 18. Particles with velocity nearly parallel to a field line at the Equator will initially move along the field line. Very soon, however, the field line curves away from the direction of particle motion. When this happens, there is a finite angle between the field and particle velocity, and the particle experiences the Lorentz force. For protons, this force is azimuthally westward, causing them to begin drifting in this direction. Now, however, there is a finite angle between the westward drift velocity and the field that creates a Lorentz force earthward. This force bends the trajectory of the particles along the field line. Together, the components of particle velocity along a field line and transverse to it cause the drift phenomenon in question.

A collection of charged particles trapped in the Earth's inner magnetic field and drifting as described above constitutes a Van Allen radiation belt. The current produced by this drift causes a magnetic field at the Earth's surface similar to that of a large ring of current in the planet's magnetic equatorial plane. Because the Earth is small compared with the size of this ring, the field is nearly uniform over the planet's surface. Its effect is to reduce the strength of the surface field. Actually, the particle drift is not confined to the equatorial plane, and the currents fill a doughnut-shaped volume defined by the shape of dipole field lines (Figure 18).

The magnetopause current. Farther still from the Earth, at about 10 R_e along the Earth–Sun line, is yet another current system that affects the surface field and profoundly changes the nature of the Earth's field in space. This system is called the magnetopause current, or Chapman-Ferraro current system after the English physicist Sydney Chapman and his student V.C.A. Ferraro who first suggested its existence. It flows in a single sheet and forms a boundary between the magnetic fields of the Earth and solar wind. When solar wind particles encounter the Earth's field, they are bent from their paths by the Lorentz force. As noted above, protons gyrate in a left-handed sense around a magnetic field and electrons in a right-handed sense. Since the particles are coming from the Sun and the direction of the Earth's field is upward parallel to its rotation axis, this gyration creates an electric current eastward in the equatorial plane as shown in Figure 19. The field of this current is such that it decreases the Earth's field outside the boundary and increases it inside. Once the current is fully developed, it occupies a thin sheet everywhere on the dayside of the Earth, outside of which is canceled all of the terrestrial field. Inside the sheet, the field is twice that of the main field.

Chapman-Ferraro current system

The magnetopause current system must close in some manner. More detailed consideration reveals that it closes on the magnetopause in much the same pattern as the dynamo currents in the ionosphere below. Figure 19 also presents a perspective view of the northern portion of the magnetopause current as seen from above the ecliptic plane. As indicated in the diagram, the current flows eastward across the dayside of the Earth and then westward around a "neutral point" (so called because the total field is nearly zero at this location). The current is symmetrical about the equatorial plane and encloses a volume of space known as the magnetosphere. Were it not for other processes (see below), the Earth's field would be completely contained inside the magnetopause. If the solar wind were absent, the field would expand indefinitely outward and produce the very simple dipole field illustrated in Figure 13.

The magnetotail current. Radially outward near local midnight rather than at local noon, there is an entirely different current system. Beginning at approximately 10 R_e and extending well beyond 200 R_e is the tail current system. This current is from dawn to dusk in the same direction as the ring current on the nightside of the Earth. In fact, it is produced by the same mechanism except that, in this region of space, curvature drift is the dominant cause of particle motion. Also, the Earth's field in this region is no longer even approximately dipolar, so the particle drift is nearly perpendicular to the Earth–Sun line rather than azimuthal around the Earth's centre. As in the case of the dayside magnetopause current, this current also closes on the magnetopause. In fact, above and below the Earth it is indistinguishable from the Chapman-Ferraro current because it closes in the same direction and is produced by the same mechanism of charge deflection. The tail current differs from the magnetopause current because over part of its path it flows interior to the Earth's magnetic field. In Figure 15 the region where this occurs is labeled the plasma sheet. For an observer on the nightside of the Earth looking away from the Sun, the current would appear to flow in a pattern similar to the Greek letter "theta." It flows westward (dawn to dusk) through the plasma sheet

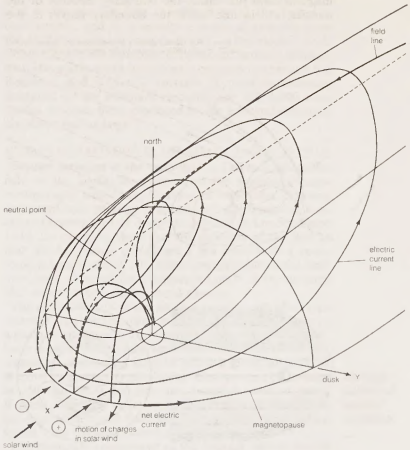


Figure 19: A perspective view of the northern portion of the magnetopause current, as seen from above the ecliptic plane. Charged particles in the solar wind are deflected in opposite directions by the Earth's main field, creating a boundary current. This current confines the field inside a finite volume called the magnetosphere (see text).

and then splits, closing above and below on the boundary of the magnetopause. Repetition of this current pattern continuously down the tail produces a current system that is essentially that of two long solenoids squashed together in a "theta" pattern, with opposite currents in the two solenoids.

Although the tail current is explained by the particle drifts discussed above, it is not obvious what process creates the tail-like magnetic field configuration required for these drifts. The Chapman-Ferraro current and the ring current are both produced in regions where the Earth's field is strong and dominated by the effects of the internal dynamo. Far from the Earth the field is stretched out into two long bundles of magnetic field lines confined by and almost wholly produced by the tail current system described above. In simplest terms, the particles travel in a field produced by their own movement. Particle motion of this type is another consequence of the interaction of the solar wind with the Earth's main field.

Effects of solar wind interaction

In the single-particle description of the solar wind interaction with the dayside magnetic field, it was noted that solar wind particles are deflected by the field and produce a current. This same interaction may be described in a fluid picture by stating that a boundary exists at a point where the magnetic pressure of the Earth's field exactly equals the perpendicular pressure of the solar wind on the boundary. On the dayside this is caused primarily by the velocity of the solar wind and not its thermal pressure.

The second component of the solar wind interaction is tangential drag, which is a frictional force exerted by the solar wind parallel to the boundary. The effect of this force is to move the Earth's field lines tailward. Two mechanisms are thought to be primarily responsible for tangential drag at the magnetopause. The first is called the viscous interaction and the second, reconnection. The latter is more difficult to visualize and will be discussed below in the section *Sources of variation in the steady magnetic field.* Both processes are summarized schematically in Figure 20.

Viscous interaction

Viscous interaction involves the transfer of momentum from the solar wind to a closed field line of the Earth's magnetic field just inside the boundary. Because of the transfer, a field line inside the boundary moves in the same direction as the solar wind. (An example of how such a transfer might occur is shown by the process of scattering a solar wind particle inside the magnetopause.)

The viscous interaction is capable of moving closed field lines from the dayside of the Earth far out on its nightside. Eventually the field lines become highly stretched into two oppositely directed bundles much like the tail of a comet except that the Earth's field is invisible. Tension in the field, combined with weakening of the tangential drag, allows the field line to return earthward. The field lines cannot return along the same path. Instead, they return through the interior of the Earth's field. The motion of these closed field lines in two closed loops, as shown in Figure 20, is called magnetospheric convection. This mechanism, together with the more important one due to reconnection, produces the tail current system.

The superposition of the Earth's main field, ring current, magnetopause current, and tail current produces a configuration of magnetic field lines quite different from that of the dipole shown in Figure 13. On the dayside the field lines are compressed inside a boundary located typically at 10 R_e. On the nightside the field is drawn out to distances probably exceeding 1,000 R_e. As will be discussed below, several processes interior to the magnetopause produce other boundaries besides the magnetopause. Several of these are evident from the Earth's surface as regions in the ionosphere within which specific types of auroras occur.

Field-aligned currents. Circulation of magnetic field lines in a pattern of closed loops within the magnetosphere is a consequence of the tangential drag of the solar wind. This circulation produces another important magnetic field source, the field-aligned current system. The field-aligned currents flow on two shells completely surrounding the Earth (Figure 21). The higher latitude shell is usually referred to as Region 1 and the lower one as Region 2. These two current sheets are caused by different physical mechanisms, but they are connected through the ionosphere and form a single circuit.

Region 1 current system

As can be seen from Figure 21, the Region 1 current originates in the region of the interface between field lines dragged tailward by the solar wind and field lines returning to the dayside of the Earth. This interface is electrically charged, positive on the dayside of the Earth and negative on the nightside. The charge on this interface is a consequence of the Lorentz force. Positive charges attached to field lines moving tailward on the dawn side of the Earth are deflected earthward toward the interface. In contrast, positive charges moving sunward just inside the interface are deflected away from the Earth (because their velocity is opposite to those on the other side of the interface). This is again toward the interface, hence a positive charge accumulates. On the dusk side the deflections are the same, but a negative charge accumulates at the interface. Because of this charge, the centres of the loops become charged like the terminals of a battery.

In the Earth's field, magnetic field lines are almost perfect conductors of current as there are no collisions to cause resistance. This allows the effects of the charge separation in the magnetosphere to be connected to the ionosphere at the feet of the charged field lines. Because the ionosphere conducts current, current can flow from the positive to negative terminals. Thus, current leaves the positive terminal of the magnetospheric "battery" and flows down field lines on the dawn side, then across the polar ionosphere, and finally out on the dusk side.

The actual current path is not nearly so simple because the ionospheric conductivity is not uniform. One source of nonuniformity is solar illumination of the dayside. Another is loss of particles from the magnetosphere to the ionosphere. This loss occurs in two rings centred around the north and south magnetic poles. Inside of these rings the ionosphere is constantly bombarded by particles that ionize the atmosphere and generate auroras. Because auroras are almost always present in these ovals, they are usually referred to as auroral ovals.

On the dayside, the particle bombardment is a result of the neutral points about which the magnetopause currents flow. These neutral points are natural funnels that allow solar wind particles to pass through the magnetopause.

From: S.W.H. Cowley, *Reviews of Geophysics and Space Physics*, vol. 20, p. 532 (1982); copyright by the American Geophysical Union

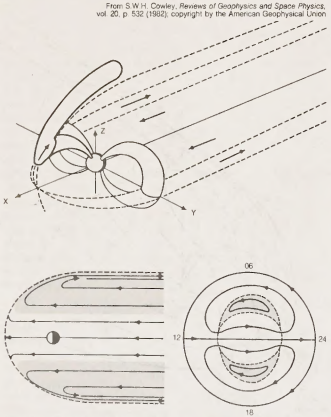


Figure 20: Two physical mechanisms create tangential drag on the outer boundary of the magnetosphere: viscous interaction and magnetic reconnection. The viscous interaction drags closed field lines along the flanks (bottom left), and magnetic reconnection transports open field lines over the poles (bottom right).

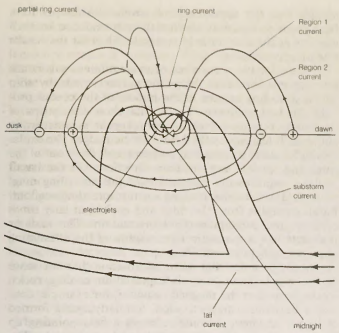


Figure 21: The field-aligned current system includes two shells of magnetic field lines connecting the magnetosphere to the ionosphere.

On the nightside the particles also originate in a natural funnel but, in this case, one produced by the projection of the plasma sheet onto the ionosphere. The particle bombardment increases the electrical conductivity of the ionosphere inside the auroral ovals relative to that in the surrounding ionosphere.

Region 2 current system

To understand the closure of the Region 1 current system, the Region 2 system must be considered. This second system is a result of charge separation by drift in the main field. As discussed in relation to the ring current, negative charges (electrons) drift eastward (in a right-handed sense) around the Earth, while positive charges (protons and heavy positive ions) drift westward. These particles preferentially approach the Earth on the nightside because of the magnetospheric convection system. As they approach the Earth, they tend to separate due to drift, with more negative charges drifting around the Earth on the dawn side and more positive charges around the dusk side. The centres of these regions also become electrically charged. Because field lines connect the regions to the ionosphere, currents can flow from them as well. In this case, the polarity is reversed from that of Region 1. Accordingly, in Region 2, current is drawn from the ionosphere on the dawn side and expelled to the magnetosphere on the dusk side.

The field-aligned current system shown in Figure 21 is a superposition of all the elements discussed above. The path of this current can be summarized as follows. Current leaves the region of interface between counterstreaming magnetic field lines on the dawn side and flows down all field lines lying in a volume connected to this region. The current then splits, some flowing across the illuminated portion of the polar cap and some flowing equatorward across the morning side of the auroral oval. The current that turns equatorward flows out along lower latitude field lines connected to the accumulation of negative charges and then flows westward across midnight as a partial ring current carried by the oppositely drifting particles. Near dusk it flows down along field lines to the ionosphere, then poleward, and finally out along field lines to the dusk interface.

At the dawn and dusk magnetopause, particles of opposite sign undergo certain actions. For example, at dawn negative charges are pushed outward toward the flowing solar wind. At dusk the opposite occurs. These charges also can discharge via field lines connected to the Earth in the region near the feet of field lines emanating from the dayside neutral points or perhaps through the solar wind by mechanisms not yet completely understood. This closure completes the electric circuit.

A surprising characteristic of the field-aligned current system is that its effects are almost completely invisible on the ground, even though it profoundly changes the field in space. Because the field-aligned current system consists of two oppositely directed, nearly parallel current sheets, its magnetic field is almost entirely confined between the sheets. The existence of this system is, however, apparent in one way. It drives a secondary ionospheric current system consisting of two convective electrojets.

Structure of the field-aligned current

Convective electrojets. The auroral electrojets are two broad sheets of electric current that flow from noon toward midnight in the northern and southern auroral ovals. The dawn side current flows westward, creating a decrease in the magnetic field on the surface. The dusk side current flows eastward and produces an increase in the magnetic field. Both currents flow at an altitude of approximately 120 kilometres in a region known as the E region of the ionosphere. In this region the collision rate between positive ions and atmospheric neutral particles is much larger than it is between electrons and neutrals. Higher in the ionosphere there are almost no collisions, while in the lower region there is little ionization. Because of the different collision rates, ions in the E region drift more slowly than electrons and thus create an electric current. At higher altitudes where equal numbers of positive and negative charges drift at the same rate, no current is produced because no net charge is transported. In the E region positive charges moving backward relative to the drift create a current opposite to the drift.

The ionospheric drift results from magnetospheric convection. Field lines with "feet" in the auroral ovals drift toward the dayside, so that the electrojet currents are toward the nightside. The electrojet currents flow at right angles to the sheets of ionospheric current connecting the field-aligned currents of Region 1 and Region 2 at the poleward and equatorward boundaries of the auroral ovals. As these currents are driven by the electric field produced by charge accumulation in the magnetosphere, they flow in the same direction as the electric field. The electrojet currents are thus at right angles to the electric field. Such a current, called a Hall current (after the Hall effect), is always present when an electric field is applied to a conductor containing a magnetic field.

Hall current

In the Earth's ionosphere the electrical conductivity parallel to the electric field is referred to as the Pedersen conductivity, and it is usually a factor of two less than the Hall conductivity perpendicular to the electric field. Consequently, the electrojet currents are actually stronger than the north–south ionospheric currents connecting the Region 1 and Region 2 currents. Typical disturbances produced by the westward electrojet are 500–1,000 nanoteslas, whereas those produced by the eastward electrojet are about half as large.

SOURCES OF VARIATION IN THE STEADY MAGNETIC FIELD

Secular variation of the main field. The main magnetic field of the Earth, as observed at the surface, changes continuously with time. Changes of very short duration compared with geologic processes are called secular variation. Observations of declination made in London since 1540, for example, show that the direction of the field at that site has nearly completed a full cycle with a peak-to-peak amplitude of 30°. Other components of the field have been observed for a shorter length of time, but they also are exhibiting similar rapid change.

The characteristics of the secular variation are often represented by superimposing maps of the rate of change of a given field component on maps of the component itself. Such maps reveal that the world may be broken down into regions of continental scale in which a given component is either increasing or decreasing. Changes can be as large as 150 nanoteslas per year and persist for tens of years. If maps of secular variation from successively later times are examined, many features of the secular variation are found to be displaced westward with time.

Manifestations of secular variation

The dominant component of the internal field is that of a centred dipole. It is useful to determine whether this component changes in the same way as the remainder of the field. Because the field of a dipole is so simple, it is more convenient to represent its change by its strength and orientation rather than by maps. Secular variation of the

non-dipole components, however, are usually presented as maps. Such maps are similar to maps of secular variation of the entire field, indicating that most of the secular change is caused by the non-dipole components. On the average, the non-dipole components of the field appear to drift westward at an average rate of 0.18° per year. At this rate, drifting features circle the Earth in only 2,000 years. Not all of the non-dipole field exhibits drift. At least half of it appears fixed and variable only in intensity.

The dipole component also changes with time. Its strength decreases with time. Since 1850 it has decreased from about 8.5×10^{22} to 8.0×10^{22} amperes per square metre. If this trend continues, the dipole component will vanish in another 2,000 years. As will be discussed in the next section, the dipole component of the Earth's field appears to be in the process of reversing.

The best estimates of the orientation of the dipole component appear to change with time. The dominant change is a westward drift of the azimuth of the dipole but at a rate much slower (0.08° per year) than the non-dipole component. The polar angles also may be increasing but even more slowly.

The origin of the secular variation is not known. Investigators suspect that it is a secondary effect of the dynamo mechanism that generates the main field. The short time scale of the variation implies that the source is in the outer region of the liquid core. If the source were deeper, the variation would be so attenuated by the electrical conductivity of the core that it would be undetectable at the surface.

Westward drift of magnetic anomalies

The westward drift of magnetic anomalies evident in the secular variation should provide an important clue to the origin of the main field if only it can be interpreted. One model explains the drift by postulating that the outer portion of the liquid core is rotating slower than the more rigid mantle above. As a whole, the Earth rotates eastward. If features within the core rotate more slowly than surface features, they will appear to move backward relative to the general rotation—*i.e.*, westward. In this model the secular variation is caused by portions of eddies in the internal current system that rotate more slowly than the planet as a whole.

A more recent model for the westward drift posits that it is produced by hydromagnetic waves in the core (see below *Magnetohydrodynamic waves—magnetic pulsations*). In this model the core rotates at the same rate as the outer mantle, but a wave propagates slowly around the outer portion of the core. Because waves in a conducting fluid distort the magnetic field frozen within it, they produce changes that can be observed at the surface. Since the characteristics of waves depend on the medium through which they propagate, it may be possible to infer properties of the outer core from surface observations.

Reversals of the main field. The Earth's internal magnetic field has not always been oriented as it is today. The direction of the dipole component reverses, on an average, about every 300,000 to 1,000,000 years. This reversal is very sudden on a geologic time scale, apparently taking about 5,000 years. The time between reversals is highly variable, sometimes occurring in less than 40,000 years and at other times remaining steady for as long as 35,000,000 years. No regularities or periodicities have yet been discovered in the pattern of reversals. A long interval of one polarity may be followed by a short interval of opposite polarity.

Available data suggest that during a reversal the strength of the dipole component shrinks to zero while maintaining its orientation. It then grows again to its former strength but with opposite orientation. During the interval in which there is no dipole component, the non-dipole part of the field appears to persist.

During field reversals the outer portion of the Earth's magnetic field is greatly altered. The absence of a dipole component would mean that the solar wind would approach much closer to the Earth. Cosmic-ray particles that are normally deflected by the Earth's field or are trapped in its outer portions would reach the surface of the planet. These particles might cause genetic damage in plant or animal communities, leading to the disappearance of one species and the appearance of another. Attempts have been made to establish whether there is evidence for such changes at the time of field reversals. Thus far the results remain inconclusive.

Evidence for magnetic field reversals

Evidence for the occurrence of magnetic reversals is unquestionable, however. Magnetic surveys made by ship across spreading centres in the middle of the oceans provide the best evidence. These data show that strips of oppositely magnetized ocean floor appear symmetrically about such features as the Mid-Atlantic Ridge. The explanation for these strips is that molten basalt flows out of the ridge and spreads away in both directions. As the basalt cools, it captures the orientation of the prevailing magnetic field and carries it along on the spreading seafloor. Basalt emerging from the ridge and cooling at later times captures the subsequent field orientation. The seafloor thus acts like a magnetic tape, capturing the alternating sequence of field orientations.

It should be noted that more information than the sense of the dipole component is captured in cooling rocks. Rocks formed at the magnetic equator, for example, contain a horizontal magnetization. Similarly, rocks formed at higher magnetic latitudes contain a field pointing up or down at an inclination that depends on latitude. The declination of the magnetization further reveals the direction to the magnetic pole at the time of the magnetization. Together, these two angles can be used to infer the location of a virtual magnetic pole relative to the location of the sample.

Such a technique has been used to study the history of the Earth's field at various locations. When virtual poles are determined from progressively older rocks, it is found that the virtual poles appear to wander with time. For many years it was thought that this "polar wandering" was a characteristic of the Earth's magnetic field. Recent studies, however, prove instead that it is a result of continental drift. Magnetic poles have not moved significantly relative to the geographic poles, but rather the continents have. Thus, progressively older rocks have been formed when continents were at different locations than they are today (see also below *Evidence for polar wandering, continental drift, and seafloor spreading*).

Cause of the reversals

Reversals in the main field must be caused by the dynamo mechanism that gives rise to the field in the first place. The time scale for the reversal is so rapid that it clearly cannot be caused by geologic processes. Furthermore, reversals cannot be caused by simple decay and reappearance of a preexisting field. The electrical conductivity of the core is too high to allow the field to decay on such a short time scale. In some way, minor changes in the magnetic field configuration of the core must be amplified by thermal convection, causing the field to grow rapidly in the opposite direction. Models that simulate the main field have been shown to possess this property. The solutions to equations that describe the generation of the main field are unstable, and small changes can cause solutions of opposite sign to appear.

Variations in the ionospheric dynamo current. The ionospheric dynamo is produced by movement of charged particles of the ionosphere across the Earth's main field. This motion is driven by the tidal effects of the Sun and the Moon and by solar heating. The ionospheric dynamo is thus controlled by two parameters: the distribution of winds and the distribution of electrical conductivity in the ionosphere. These parameters are influenced by several factors, including the orbital parameters of the Earth, Moon, and Sun; the solar cycle; solar flares; and solar eclipses. Changes in the position of the Sun and the Moon relative to the Earth as a result of orbital motions cause variations in distance. This alters the strength of the tides and of solar heating, thereby changing ionospheric wind patterns. These changes are apparent as a seasonal modulation of the winds and hence of the strength of the current.

The second parameter that controls the dynamo current is the electrical conductivity of the ionosphere. Any process that alters ionospheric conductivity changes the current. On the dayside of the Earth, the dominant source of ionization is sunlight. The amount of ionization depends

on the angle at which sunlight enters the atmosphere. Vertical incidence produces more ionization per unit volume than slant entry. For a given hemisphere, normal incidence occurs in summer. Thus, this effect also causes a strong seasonal modulation of the dynamo current.

Impact of sunspots and other solar activity

The degree of atmospheric ionization also depends on the phase of the solar cycle. This 11-year cycle of sunspot activity produces variations in the amount of ultraviolet radiation emitted by the Sun. More sunspots lead to more ultraviolet radiation and increased ionospheric conductivity, hence stronger currents. On a shorter time scale solar flares emit X rays that penetrate deeper in the atmosphere, temporarily ionizing the D region. Dynamo currents are then produced in this layer by whatever winds are present there.

A solar eclipse produces the opposite effect on ionospheric conductivity. The shadow of the Moon as it crosses the ionosphere decreases ionization. Recombination of ionospheric electrons and ions in the absence of light quickly reduces the conductivity. Because the effect is localized and of short duration, its effect on the overall dynamo current is slight.

Magnetic storms—growth of the ring current. The ring current is produced by the drift around the Earth of charged particles of the outer Van Allen radiation belt. During quiet conditions the effect of this current at the Earth's surface is negligible (~20 nanoteslas). Once or twice a month there occurs a phenomenon known as a magnetic storm, during which the intensity of the ring current increases and produces disturbances that are typically on the order of 100 nanoteslas but that can be as large as 500 nanoteslas. A variety of phenomena that affect humans occurs during magnetic storms. A few of these include increased radiation doses for occupants of transpolar flights, distortion of compass readings in polar regions, disruption of shortwave radio communications, increased corrosion in long pipelines, failure of electrical transmission lines, anomalies in the operations of communications satellites, and potentially lethal doses of radiation for astronauts in interplanetary spacecraft. Efforts have been undertaken to mitigate such serious problems. In the United States, for example, the federal government operates a Space Disturbance Forecast Center in Boulder, Colo., which monitors the state of the Sun and solar wind and attempts to predict the occurrence of such "space weather."

Cause of magnetic storms

It is known that magnetic storms are produced by a change in the properties of the solar wind. Magnetically quiet times occur when the solar wind contains a magnetic field called the interplanetary magnetic field (IMF) that has the same direction as the Earth's field on the dayside. Magnetic disturbances occur when this field rotates toward an antiparallel orientation. Normally, the IMF lies in the ecliptic plane, which on the average is roughly parallel to the Earth's magnetic equator. Small departures from this average orientation are caused by rotation of the tilted dipole magnetic field once per day and by revolution of the Earth around the Sun once per year. Large departures are caused by changes in the direction of the IMF relative to the ecliptic. Such changes are produced by several phenomena that originate on the Sun.

The most spectacular event that may cause a magnetic storm is a solar flare, which is an explosion in the corona of the Sun that releases an enormous amount of energy in the form of outward-streaming particles. The bulk of these particles takes approximately two days to arrive at the Earth, where they begin to influence its magnetic field. During transit the solar flare particles catch up with slower particles emitted earlier. The subsequent interaction of the high- and low-speed solar wind components causes a high-pressure region to develop, and this region tilts the IMF out of the plane of the ecliptic. If the IMF is tilted antiparallel to the Earth's field, a magnetic storm results.

Another phenomenon responsible for magnetic storms is the existence of coronal holes around the Sun. X-ray images of the Sun made during the 1970s by the U.S. Skylab astronauts revealed that the corona of the Sun is not homogeneous but often exhibits "holes"—regions within the solar atmosphere in which the density of gas is lower than in adjacent regions and from which charged particles escape with relative ease. Particles from such holes reach higher velocities in their outward expansion than do normal solar wind particles and produce high-speed streams. These streams interact with the slower speed solar wind emitted from regions without holes and produce the same tilting of the IMF described above. Coronal holes persist for many 27-day solar (equatorial) rotations and, as a consequence, produce recurrent magnetic storms. Coronal holes are the hypothetical m regions proposed many decades ago to explain recurrent storms that could not be associated with particular solar flares.

The observed dependence of geomagnetic activity on the orientation of the IMF is explained by most researchers as a consequence of magnetic reconnection. In reconnection, pictured schematically in Figure 22, two oppositely directed magnetic fields are brought together by flowing plasmas at an x-type neutral line. Far from the neutral line the magnetic field is frozen in the plasma; however, near the neutral line it becomes unfrozen and diffuses through the plasma, establishing a new configuration of magnetic field lines. On passing through the neutral line, field lines from opposite sides connect and flow rapidly away from the neutral line at right angles to their direction of inflow. In the process, energy originally stored in a strong magnetic field is converted to the kinetic energy of flowing plasma. In addition, the topology of magnetic field lines is changed. At the dayside magnetopause (see Figure 15),

Magnetic reconnection

From V.M. Vasyliunas, *Reviews of Geophysics and Space Physics*, vol. 13, p. 311 (1975); copyright by the American Geophysical Union

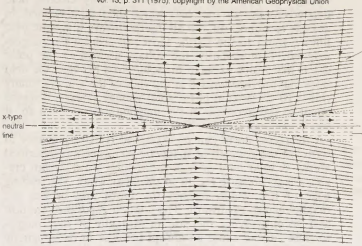


Figure 22: Magnetic reconnection at an x-type neutral line allows two plasmas threaded by magnetic fields to flow together and merge. (Plasma flow is represented by the dashed lines and the magnetic field lines by the solid lines.) A strong magnetic field in the inflow is converted to particle kinetic and thermal energy in the outflow region.

field lines of the IMF become connected to geomagnetic field lines. Because the IMF is frozen into the solar wind, the portion of the reconnected field line external to the magnetosphere is dragged away from the Sun above and below the polar caps. The portions of the field line inside must follow the external portions and, hence, their "feet" appear to drift across the polar caps. This process cannot go on indefinitely as geomagnetic field lines will be continuously eroded from the dayside unless they are replaced by an internal flow. Such a flow develops after a short lag and follows the same pattern as the return of field lines drawn away from the Sun by viscous interaction. When the flow is fully developed, the flux of magnetic field lines toward the Sun within the magnetosphere balances the flux away from the Sun above and below the polar caps.

For field lines to return from the nightside, they must first disconnect from the solar wind. This occurs at a second x-type neutral line located behind the Earth (see Figure 15). There, as on the dayside, oppositely directed field lines are brought together by plasma flows. Reconnection occurs, and the IMF and geomagnetic field lines again become separate entities.

The topology of magnetic field lines produced by the reconnection process accounts for the existence of auroral ovals. As evident from Figure 15, field lines of the polar caps are "open" to the solar wind, whereas those at lower latitudes are "closed." On the nightside the field lines con-

necting to the neutral line form a natural boundary for trapping charged particles. The region interior to the "last-closed field lines" is filled with trapped particles and is called the plasma sheet. The projection of the last-closed field lines on the polar atmosphere forms the poleward boundary of the nightside auroral oval. As previously noted, a second boundary forms on the nightside of the Earth as particles drift earthward under the influence of magnetospheric convection (driven by both viscous interaction and reconnection) and then enter the region of strong azimuthal drift. This boundary is called the inner edge of the plasma sheet, and it projects as the equatorward edge of the nightside auroral oval.

Generation of a magnetospheric electric field

An important consequence of reconnection is that it produces a magnetospheric electric field, as does viscous interaction. This comes about as a result of the connection between the interplanetary and geomagnetic fields. This process can be understood as follows. In the solar wind the Lorentz force separates positive and negative charges, just as it does in the magnetospheric boundary layer. These charges accumulate at boundaries within the solar wind where either the velocity or orientation of the IMF changes. There is an electric field between these boundaries. Because magnetic field lines have nearly infinite conductivity, the electric field originating in the solar wind is projected by magnetic field lines into the magnetosphere and onto the polar caps. The effect of this field depends on its strength and the length of the dayside x line. The voltage, or potential, drop caused by any electric field depends on the distance over which the field is applied. In dayside reconnection, not all interplanetary magnetic field lines connect to the Earth. Most slip around the magnetosphere. Consequently, the voltage applied to the polar cap is that which exists in the solar wind between the field lines that are reconnected at the ends of the x line. Usually this is 10–20 percent of the total voltage across a distance equal to the diameter of the magnetosphere. Even so, it can be as large as 200,000 volts.

A magnetic storm can be explained relatively simply in terms of the concept of magnetic reconnection described above. A solar flare or high-speed solar wind stream creates a high-pressure region in the solar wind. The leading edge of this region reaches the Earth and presses the magnetopause earthward. The sudden earthward motion and accompanying increase in strength of the magnetopause current cause an abrupt increase in the magnetic field at the Earth's surface known as the storm sudden commencement. In most cases, the pressure remains high for a number of hours and causes a larger than normal surface field. This interval is called the initial phase of a magnetic storm. Eventually, the IMF turns toward the south, antiparallel to the Earth's field, and magnetic reconnection begins. Closed magnetic field lines are eroded from the dayside and added to the polar caps, increasing their diameter. The aurora, which occurs in two ovals immediately equatorward of the polar caps, moves to lower latitudes. Within about an hour the nightside neutral line begins to return a sufficient amount of flux to the dayside, and convection approaches equilibrium.

Principal driving mechanism of magnetospheric convection

Magnetic reconnection drives magnetospheric convection much more efficiently than does viscous interaction. Consequently, all phenomena associated with convection are much enhanced over quiet times. Convecting particles approach closer to the Earth before they are deflected by drift in the main field. Field-aligned currents and the ionospheric electrojets driven by the convection electric field are much stronger. In addition, particles drifting across the main field gain more energy. This process of energization occurs at all times but is much enhanced during strong convection. It is caused by the dawn-to-dusk electric field across the magnetosphere. Any positive charge that drifts in the direction of an electric field gains energy from the field. Since positive charges on the nightside drift toward dusk, they gain energy. Similarly, electrons gain energy drifting toward dawn opposite to the electric field. On the dayside the drifts are reversed and particles lose energy. The combination of effects from more particles drifting faster closer to the Earth enhances the nightside ring current and reduces the magnetic field on the Earth's surface.

If the magnetospheric electric field remained steady, the particles drifting around the Earth would lose their energy on the dayside and convect to the magnetopause, where they would be lost to the solar wind. If the electric field across the magnetosphere is suddenly reduced by a northward turning of the IMF, however, many particles that would have been returned to the solar wind by convection are trapped on drift paths closed around the Earth. These particles rapidly separate into a doughnut-shaped ring that forms a symmetrical ring of current around the planet. Subsequent cycles of increase and decrease in the magnetospheric electric field trap additional particles and increase the energy of those already trapped. By this and another process described below, the ring current grows and produces the main phase of a magnetic storm.

A second and more spectacular phenomenon also contributes to the development of the storm main phase. This phenomenon is known as a magnetospheric substorm. The term substorm is used because such an event is observed during the development of the main phase of a storm. Since events of this kind occur more frequently at times when there is no significant growth of the ring current, they are treated below as a separate topic. As will be shown, the main effect of a substorm is energization and injection of particles into the inner magnetosphere in a localized region near midnight. Although the particles do not appear to have an immediate effect on the strength of the ring current, they are usually trapped on closed drift paths and are available for subsequent energization by fluctuations in the magnetospheric electric field. Many of the dramatic and often detrimental effects attributed to magnetic storms are actually caused by particularly intense substorms that accompany them. Both phenomena are linked by the same fundamental processes (see below).

The particles of the ring current have a finite lifetime before being lost to the Earth's atmosphere. Two processes—charge exchange and wave-particle interactions—contribute to this loss. Charge exchange is a process wherein a cold atmospheric neutral particle interacts with a positive ion of the ring current and exchanges an electron. The ion is converted to an energetic neutral, which, since it is no longer guided by the main field, may be lost in the deeper atmosphere, exchange again with an ion farther from the Earth, or be lost from the magnetosphere entirely. The previously neutral particle becomes charged in this process and is subsequently subject to drift in the main field, albeit with lower energy than the original ion. This process of charge exchange is dependent on the number of particles present in the ring current. As the number increases, so does the rate of decay due to charge exchange. For any given rate of injection into the ring current, the current grows until the rate of decay balances the rate of injection. At this point, the ring current becomes stable and persists as long as steady injection continues.

Charge exchange and wave-particle interactions

In a typical magnetic storm the interval during which the IMF is tilted out of the ecliptic antiparallel to the Earth's main field is on the order of eight to 16 hours. The lifetime of a particle against charge exchange is about the same. Accordingly, it is rare that equilibrium of the ring current ever develops. Instead, the IMF turns northward and the ring current gradually decays. In most cases, this recovery phase of the magnetic storm lasts for two to three days before quiet conditions are reestablished.

A second process that contributes to the decay of the ring current is the cyclotron instability of particles gyrating in the Earth's field. In this process an electromagnetic wave with a frequency near that at which particles gyrate about the field interacts with the particles exchanging energy. If conditions are right, the wave gains energy at the expense of the particle and in the process scatters the particle, so that it tends to follow a field line more closely. A succession of such scatterings eventually produces a particle moving directly along a magnetic field line. The particle then travels all the way to the atmosphere and is lost from the ring current. The appropriate condition for this process occurs when the ring current possesses more particles near the equatorial plane than near the end of the field line. Magnetospheric convection produces this situation in the inner magnetosphere; thus, this process is an

important loss mechanism contributing to the observed ring-current decay. In a typical ring current the waves produced by protons have a frequency between 0.2 and five hertz. Electrons produce waves of about 1,836 times higher frequency.

Magnetospheric substorms—unbalanced flux transfer. Magnetospheric substorm is the name applied to the collection of processes that occur throughout the magnetosphere at the time of an auroral and magnetic disturbance. The term substorm was originally used to signify that the processes produce an event, localized in time and space, which is distinct from a magnetic storm. During a typical three-hour substorm, the aurora near midnight exhibits a sequence of changes called the auroral substorm. Accompanying the changes in the aurora is a sequence of magnetic variations referred to as the polar magnetic substorm. Most of the detrimental effects of a magnetic storm are caused by the substorms that accompany them.

An isolated substorm begins when the IMF turns southward and dayside reconnection begins. For about an hour afterward, bands of quiet auroral arcs drift equatorward near midnight in the northern and southern auroral ovals. The eastward and westward electrojets, flowing from noon toward midnight along the ovals, gradually increase in strength and move equatorward along with the aurora. This quiescent phase is called the growth phase of the substorm.

The growth phase is terminated by a sudden brightening and activation of the most equatorward arc in each oval. This event is often termed the auroral breakup, and it signals the onset of the substorm expansion phase. Soon after onset, auroral activity expands to fill the entire sky above a particular ground observer. Rapid motion, development of vertical rays and folds, and the appearance of colour at the bottom of auroral forms are characteristic features of this phase. Detailed observations made from the ground and images from satellites reveal that the region of auroral disturbance expands poleward and westward. A surge of bright aurora, known as the westward traveling surge, propagates to the west and eventually decays into drifting bands that sometimes pass the dusk meridian. On the dawn side, patches of pulsating aurora and large omega-shaped bands drift eastward.

Auroral breakup

Accompanying the aurora are simultaneous changes in the magnetic disturbances. The most important of these is an enhancement of the westward electrojet in the region of the expanding aurora. As the surge travels westward, so too does the leading edge of the enhanced electrojet. On the ground the magnetic field suddenly decreases, sometimes by as much as 2,000 nanoteslas as the surge passes overhead. Behind the advancing fronts of the aurora, the particles responsible for the auroral light also increase the electrical conductivity of the ionosphere and cause the convection electrojets to increase in strength. The expansion phase of the substorm terminates after about 30 minutes, and the final phase begins.

The final phase of a substorm is called the recovery phase. During this phase the aurora and currents gradually drift back to their original equatorward locations as they simultaneously decrease in luminosity and strength. Provided that the IMF has turned northward in the intervening time, the recovery phase ends after approximately 90 minutes.

Often the IMF does not turn northward immediately; it may fluctuate between north and south. In such cases, the auroral and magnetic disturbances become much more complex and are not easily characterized. Situations of this kind usually persist for a sufficient length of time, so that many particles are brought into the inner magnetosphere where they are energized and trapped and produce a magnetic storm. Nonetheless, many features of the isolated substorm can still be recognized.

Substorms as manifestations of time-varying convection

The magnetospheric substorm also can be explained in terms of magnetic convection driven by magnetic reconnection. A substorm, however, is a manifestation of time-varying convection. In the reconnection model of substorms, transport of magnetic flux and particles never reaches equilibrium. During the growth phase of a substorm, magnetic flux is eroded from the dayside and added

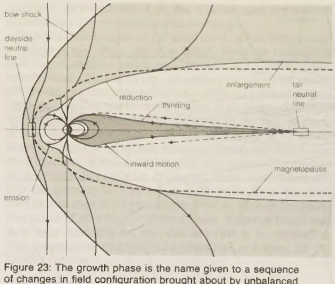


Figure 23: The growth phase is the name given to a sequence of changes in field configuration brought about by unbalanced flux transfer.

to the lobes of the magnetotail. As illustrated in Figure 23, the dayside magnetopause moves inward as a result of the flux lost, while the polar caps increase in size as a result of the flux gained. The additional flux in the near-tail requires an increase in the tail field and hence in the tail current, since the additional flux is contained in a volume of smaller cross section than was the initial quiet-time flux. Also, because the tangential drag on the tail has increased, the tail current moves earthward to increase the force that the Earth exerts on the tail, thus balancing the additional force of the solar wind. Closed flux simultaneously begins returning to the dayside and emptying the nightside plasma sheet. Equatorward motion of the aurora during this phase is simply a manifestation of the increasing size of the tail lobes. Enhancements of the eastward and westward electrojets are a consequence of the increased rate of convection driven by the southward IMF.

The expansion phase of a substorm

The expansion phase is less well understood than the growth phase. Many investigators support the "near-Earth neutral-line" model, but concurrently other explanations have been suggested. In the neutral-line model a localized x-type neutral line is formed inside the plasma sheet somewhere between 20 and 40 R_e (earth radii) behind the Earth. Figure 24 (top) shows the topology of the magnetic field when such a line is first formed. In the noon–midnight meridian of the magnetotail the magnetic field is divided into several regions by the simultaneous presence of two x-type neutral lines. Between the two x lines is an o-type neutral line around which there are closed loops of magnetic field. This field connects to neither the solar wind nor the Earth and remains in place only because it is surrounded by a sheath of field lines attached to the Earth. This geometry persists only as long as the sheath remains. Eventually reconnection severs the last-closed field lines, and subsequently open field lines of the tail lobe begin to reconnect. Shortly after this happens, the region of closed field lines is sheathed by field lines connected to the solar wind, as shown in Figure 24 (bottom). Tension in these field lines pulls the bubble of plasma and field, or plasmoid, as it is called, from the centre of the magnetotail. The plasmoid travels down the tail, collapsing the plasma sheet behind it.

In the neutral-line model, the sudden brightening of the auroral arc near midnight is thought to occur when reconnection reaches the last-closed field lines. The subsequent poleward expansion of the aurora is interpreted as the boundary of lobe field lines moving into the near-Earth neutral line to be reconnected. Finally, the westward surge is explained as an expansion of the azimuthal extent of the near-Earth neutral line by some as yet unexplained process.

In this model the final recovery stage of an isolated substorm is produced by a rapid tailward motion of the near-Earth neutral line. This probably occurs when there is no longer excess magnetic flux in the tail lobes to be returned to the dayside. Once this happens the magnetic field and plasma flow in the near-Earth region of the tail return

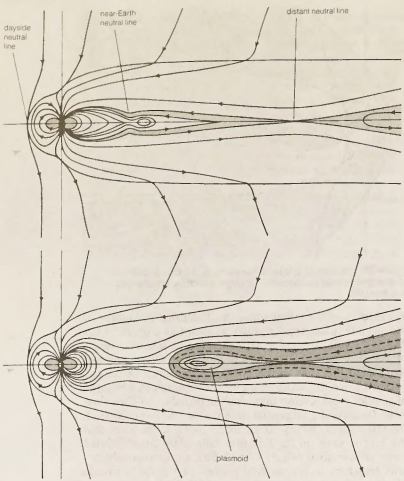


Figure 24: (Top) Magnetic reconnection inside the closed magnetic field lines of the plasma sheet, which produces a bubble of plasma and field called a plasmoid. The plasmoid is initially held in place by closed field lines attached to the Earth. (Bottom) When field lines in the boundary of the plasma sheet become reconnected, the plasmoid is pulled from the Earth's magnetotail by field lines connected to the solar wind.

to quiet-time conditions and reestablish the pre-substorm conditions of aurora and magnetic disturbance.

An essential feature of this model is that the near-Earth neutral line is azimuthally localized. To achieve this localization, it is necessary to divert a portion of the tail current to the ionosphere at the ends of the neutral line. The sense of this diversion is downward toward dawn and upward toward dusk, as shown schematically in Figure 25. In the ionosphere the current flows westward and enhances the preexisting westward convection electrojet. This current system is called the substorm wedge and, though not illustrated, connects symmetrically to both northern and southern auroral ovals.

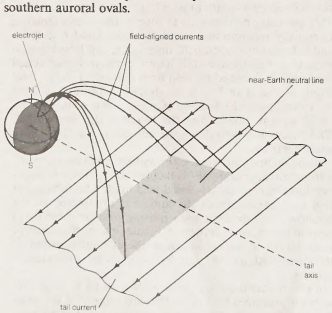


Figure 25: The substorm-wedge current is a field-aligned current system created when a localized x-type neutral line forms close to the Earth. A portion of the tail current is temporarily diverted through the ionosphere.

Substorm-wedge current system

The substorm-wedge current system causes sudden changes in the magnetic field at the Earth's surface during substorms. These changes induce very strong localized electric fields. These transient electric fields energize particles to high energy and propel them earthward. Loss of these particles to the atmosphere causes the aurora within the expanding bulge of the auroral substorm and later, as they drift, the ionization of the atmosphere that enhances electrical conductivity. Many particles also are trapped in drift paths around the Earth, adding to the particles in the ring current. On the ground, the same induction effects are responsible for the disruption of electrical transmission lines and for corrosion in pipelines. Changes in radio propagation are caused both by the changing size of the polar cap relative to lower latitude regions and by increased absorption of radio waves in the ionization occurring at the bottom of the ionosphere.

Magnetohydrodynamic waves—magnetic pulsations. A major source of variations in the Earth's magnetic field is magnetohydrodynamic waves. These waves originate in the outer magnetic field and propagate along field lines to the Earth's surface. On reaching the surface they cause minute oscillations in the magnetic field (hence their older name, micropulsations). These waves typically have amplitudes ranging from 100 to 0.1 nanoteslas, with lower frequencies exhibiting larger amplitudes. Magnetic pulsations have been classified phenomenologically on the basis of waveform into pulsations continuous (Pc) and pulsations irregular (Pi). Each class is subdivided into different frequency bands supposedly on the basis of boundaries defined by different generation mechanisms. By definition, magnetic pulsations fall into the class of electromagnetic waves called ultra low frequency (ULF) waves, with frequencies from one to 1,000 megahertz. Because the frequencies are so low, the waves are usually characterized by their period of oscillation (one to 1,000 seconds) rather than by frequency.

Magnetic pulsations

Until recently little was known about the causes of these waves. Improvements in instrumentation, notably DC amplifiers and spacecraft-borne devices, however, have contributed significantly to their understanding. There is a variety of mechanisms that produce such waves. The simplest mechanism is perhaps the resonant oscillation of the Earth's main magnetic field in response to waves in the solar wind. In this process a broad spectrum of waves of different frequencies is generated by some process in the solar wind. A small fraction of the energy in these waves penetrates the magnetopause. Within the magnetosphere each magnetic field line has a characteristic frequency of oscillation determined by its length, the strength of the field along it, and the mass of the particles attached to it. If the waves entering the magnetosphere have the same frequency as the field line, they force it to oscillate. If there is little damping of the oscillation, its amplitude may grow large enough to be observed at the ends of the field line. Additional sources of excitation include waves on the magnetopause stimulated by flow of the solar wind, sudden pressure pulses that move the magnetopause in or out, and sudden changes in the flow direction of the solar wind that cause the magnetotail to flap.

Generation of magnetohydrodynamic waves by cyclotron instability

Another type of generation mechanism is the cyclotron instability mentioned earlier in the discussion of ring-current decay. This mechanism illustrates the way in which a plasma may lower its total energy by creating waves. In this mechanism, a wave traveling along a field line interacts with a gyrating particle on the same field line. For energy to be exchanged, the electric field of the wave must rotate with the same frequency as that of the gyrating particle. If the particle has parallel as well as gyrational velocity, it is the wave frequency Doppler shifted to the frame of reference of the moving particle that is important.

Other instabilities are related to different periodicities in particle motion. Typical examples are bounce resonance of waves with particles traveling along field lines, or drift resonance with particles drifting around the Earth. In either case the electric field of the wave and the velocity of the particle must remain in phase with each other for a significant time so that energy is exchanged. (R.L.M.)

The structure and composition of the solid Earth

The basic structure and composition of the Earth's interior have been known since the mid-20th century. In a landmark paper published in 1952, the American geophysicist Francis Birch described the constitution of the planetary interior based on a broad array of seismological, experimental, and geochemical observations. Although there have been numerous advances in the intervening years, such developments have served largely to reinforce or extend the picture described by Birch. Thus, it is worth summarizing this basic picture before describing the detailed evidence on which current models of the Earth's interior are based.

Regions of the Earth

The Earth consists of two major regions: a central core, which is almost completely molten, surrounded by a predominantly solid shell comprising the mantle and crust together (Figure 26). The chemical compositions of these two regions are entirely different. The core is made of a dense, iron-rich metallic alloy, in contrast with the outer shell that consists of rocky (or ceramic-like) material. Oxides—specifically silicate minerals (compounds of silicon and oxygen)—make up this rocky material, with the crust containing somewhat more silicon, aluminum, and calcium relative to the mantle (Table 3). Examples of such minerals that are common in the outer few hundred kilometres of the planet (roughly down to 400 kilometres beneath the surface) include olivine, pyroxene, and garnet.

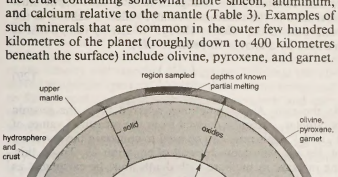


Figure 26: Schematic cross section illustrating the shell structure of the Earth.

Pressures and temperatures increase with depth inside the Earth, reaching maximum values of 364 gigapascals (GPa; 3,640,000 atmospheres) and about 6,000 kelvins (K; 10,300° Fahrenheit) at the centre. The interior temperatures are high enough to partially melt a small fraction of the crust and mantle and to completely melt the outer core. Most of the interior, however, including the inner core, is at a temperature below the melting point.

At depths of about 300 to 700 kilometres, pressures and temperatures become sufficiently high that the minerals of the upper mantle transform to more tightly packed crystal structures such as that of perovskite (see below). Because of the occurrence of these pressure-induced transformations, such physical properties as density and elastic-wave velocities are observed to increase rapidly with depth. This depth interval, called the transition zone, occurs about one-fifth of the way down into the mantle, separating the upper mantle from the lower mantle (Figure 26). Few, if any, significant mineral transformations seem to occur in the lower mantle, which is the single largest uniform region of the interior.

The fluid nature of the outer core has one consequence that can be felt directly at the surface. Turbulent flow of the liquid metal in the core effectively produces a dynamo. As noted above, the result is the main geomagnetic field (in Figure 26 the magnetic field and the fluid flow by which it is created are schematically indicated by $\boldsymbol{H}$ and $\boldsymbol{\vec{u}}$, respectively).

Although the mantle is solid in the conventional sense that its temperature is almost entirely below the melting point, this crystalline region is found to behave like a fluid over geologic time. Large-scale deformation of the mantle results in plate tectonics at the surface and the related phenomena of earthquakes and volcanoes.

This dynamic picture of the solid interior has been recognized since the 1960s, deriving largely from the realization that the crystalline solids making up the planet are ultimately weak. That is to say, laboratory experiments reveal that rocks can undergo substantial deformations over geologic time scales, particularly at the high temperatures of the interior. Also, innumerable geologic and geophysical observations at the Earth's surface are found to fit neatly into this picture of global tectonics involving large-scale deformations of the solid interior (see below *The surface of the Earth as a mosaic of plates*). Thus, it is concluded that the primary way by which heat is lost from the deep mantle and core is by convective flow, as though the mantle consisted of a highly viscous fluid. Viewing the planet as a large heat engine, one can see that it is convection that governs the thermal and chemical evolution of the Earth's interior.

ZONAL STRUCTURE AS REFLECTED BY VARIATIONS IN PHYSICAL PROPERTIES

Seismology: wave-velocity and density distributions. The main difficulty with studying the Earth's interior is apparent from Figure 26. Rock fragments (called xenoliths) are brought up volcanically from depth, thus providing samples of the upper mantle. These samples, however, seem to originate at depths no greater than 150 to 200 kilometres. Therefore, the material making up more than 90 percent of the interior is inaccessible to direct observation, and investigators must instead turn to indirect, geophysical approaches. Of these, seismology offers by far the most precise and detailed picture of the interior.

Seismic waves

Seismic waves are essentially elastic deformations that are generated by earthquakes, natural explosions (*e.g.*, volcanic eruptions), or artificial explosions (either nuclear or large chemical-explosive blasts). The waves propagate around or through the Earth, thereby revealing its internal structure.

Three types of seismic waves are commonly identified: body waves, free oscillations, and surface waves. Body waves are those passing through the interior, as illustrated in Figure 27. For large earthquakes, standing waves of the entire planet—free oscillations—are also excited. Intermediate between these types of seismic waves are surface waves, which travel around the surface of the globe. The definitions of these three wave types overlap: for example, free oscillations can be thought of as surface waves that are appropriately superposed to give a standing-wave pattern.

In general, seismic waves are similar to sound waves except that the periods of oscillation are far longer. Periods shorter than 0.1 second (frequencies of oscillation higher than 10 hertz), for example, are rarely considered in body-wave seismology, and this is at the lower limit of audible sound. Even though shorter-period (higher-frequency) waves are generated at the earthquake source, these are rapidly attenuated as they propagate through the Earth.

The longest period oscillations considered in seismology are those of free oscillations. These can range up to a period of 54 minutes (frequency of 0.3 millihertz). The limitation in this case is the size of the Earth: longer periods would require wavelengths longer than can be accommodated across the planet. The periods of surface waves are generally intermediate between those of body waves and free oscillations.

Two kinds of body waves are propagated through the solid parts of the Earth. They are longitudinal and transverse waves, which differ in the sense of deformation (or particle motion) associated with the wave. In geophysical literature they are commonly referred to as compressional (or *P*) and shear (or *S*) waves, respectively.

Compressional and shear waves

A compressional wave is in fact identical to a normal sound wave in that the deformation is in the same direction as that in which the wave is propagating. In contrast, the deformation in shear waves is perpendicular to the direction of propagation. As a result, shear waves can be transmitted only through solids. Because fluids have no rigidity (by definition), they cannot support the transverse deformation associated with shear waves.

Formally, the relationship between seismic-wave velocities and elastic properties is given by

$$V_P = \sqrt{\frac{K_S + \frac{4}{3}\mu}{\rho}}, \tag{27}$$

$$V_S = \sqrt{\frac{\mu}{\rho}}, \tag{28}$$

for compressional and shear waves, respectively. Both velocities depend on the rigidity (μ) and density (ρ), but V_P depends on the incompressibility (K_S) as well. Thus, in a fluid $\mu = 0$, $V_P = \sqrt{K_S/\rho}$, and $V_S = 0$. These relations, which are exact for a material that is isotropic in its elastic properties (K_S, μ), must be modified slightly in the case of anisotropy (*i.e.*, in case velocities or elastic properties vary with direction). For purposes of the present discussion, anisotropy can be ignored.

If the velocities inside the Earth were constant, seismic waves would travel in straight lines. Instead, the waves are reflected and refracted because of the variations in velocity, as is shown in Figure 27. These effects are completely analogous to the reflection and refraction of light waves caused by variations in refractive indices within a medium or between two mediums. (The index of refraction is simply the velocity of light in a material divided by the velocity in a vacuum.) In this sense, the body-wave seismologist treats the Earth as a large distorting lens and uses the distortions of the seismic ray paths to determine the variations of elastic-wave velocities throughout the interior.

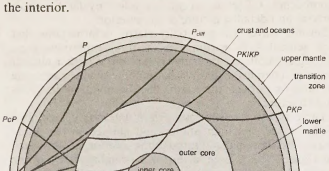


Figure 27: Cross section with shading proportional to the velocities of compressional (*P*) waves through the Earth. Several ray paths for compressional body waves are shown with labeling conventional in seismology: *P*, *K*, and *I* for compressional-wave paths outside the core, in the outer core, and in the inner core, respectively; *c* for a reflection from the core–mantle boundary; and subscript $_{diff}$ for a wave diffracted from the core–mantle boundary. Many more ray paths are observed, as are analogous paths for shear (*S*) waves.

Ray paths of body waves

A number of compressional-wave paths are illustrated in Figure 27. Many more paths have been observed, as have the analogous ray paths for shear waves. Classical seismology is based on measuring the time required for *P* and *S* waves to arrive at a seismic-wave recorder at the surface (seismometer) after following one of these paths from the earthquake source. For a given earthquake, several arrivals are usually observed at each seismometer; these correspond to waves arriving along different ray paths. Thus, based on studies of a large number of earthquakes, it is possible to infer the distribution of compressional- and shear-wave velocities throughout the Earth's interior.

Just such an analysis led to the travel-time tables and velocity-depth profiles published by Harold Jeffreys, the aforementioned British geophysicist, and Keith E. Bullen, a New Zealand seismologist, in the mid-1930s. Although there were several earlier and contemporary studies of travel times (by the American seismologists Beno Gutenberg and Charles F. Richter, for example), and numerous studies since then, the Jeffreys–Bullen model is considered to be an excellent first approximation for the compressional- and shear-wave velocity distributions with depth throughout the Earth.

In addition to the measurements of travel times, the determination of velocity distributions by body waves has been greatly enhanced since the late 1960s by the quantitative study of waveforms—the actual ground displacements as a function of time at each seismometer. The detailed character of these displacements, which might typically amount to a few to tens of micrometres (one micrometre equals 0.001 millimetre or roughly 0.000039 inch) for distant earthquakes, is extremely sensitive to the variation of velocity with depth. Therefore, if one starts with a reasonable estimate of the velocity-depth profiles (*e.g.*, based on travel times), these can be refined by calculating the waveforms from the velocity distributions and modifying these distributions in order to obtain the best match with the observed waveforms.

Study of waveforms

One of the problems with body-wave seismology is that only the V_P and V_S distributions can be determined; the individual variations of density and elastic moduli cannot be separated (equations 27 and 28). As a result, Francis Birch had to infer a density distribution for the interior in a roundabout way in his 1952 paper. First, any acceptable profile of density with depth must match the known density and moment of inertia for the Earth. Second, for a region of the interior that is uniform with depth (meaning that the mineral phases and overall composition remain constant), the change in density with depth must be related to the incompressibility, which is defined as the reciprocal of the relative change in density ($d\rho/\rho$) that is induced by a change in pressure (dP):

$$K_S = \left(\frac{dP}{d\rho/\rho}\right)_s. \tag{29}$$

As the ratio K_S/ρ (sometimes referred to as the seismic parameter, φ) can be derived from the measured values of V_P and V_S, this ratio can be used to constrain the variation of density with depth by way of equation 29. To do this, the increase in pressure with depth must be calculated as outlined below (equation 31).

Regions across which velocities vary smoothly, such as the lower mantle or the outer core, can plausibly be thought of as uniform with depth (see Figure 28). It is therefore reasonable to relate the density-depth profiles and the velocity distributions (specifically, $\varphi = V_P^2 - \frac{4}{3}V_S^2$) through these smooth regions. Once density is known with depth, profiles of the individual elastic moduli K_S and μ can be derived from V_P and V_S. This is an important accomplishment because density and elastic moduli vary in unrelated ways from one material to another or with changes in pressure and temperature; therefore, the velocities vary less systematically than either the moduli or density by themselves. It is the variation of these three properties, ρ, K_S, and μ, that are primarily used for inferring the nature of the Earth's interior.

The problem of deriving density profiles in this roundabout way from body-wave measurements was superseded

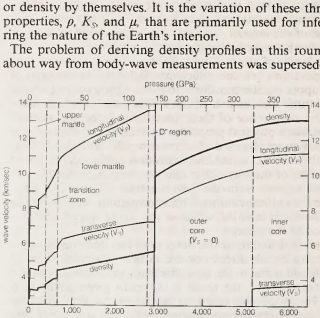


Figure 28: Summary of the average seismic-wave velocity and density profiles through the Earth according to the PREM model. The velocities of compressional (V_P) and shear (V_S) waves are given on the left, density on the right, and pressure as a function of depth on the top scale (see Table 2).

as of 1960. The first observations of free oscillations were made that year, and these allowed for determinations of the density variation throughout the interior.

Large earthquakes set the globe to vibrating in numerous modes of free oscillation. Dozens of these overtones have been recognized, and they can keep the Earth "ringing" for days after an earthquake. Two types of modes are observed: spheroidal and toroidal. These types involve motions toward (or away from) the centre of the globe and torsional motions of the globe, respectively. Evidently, these are roughly analogous to the compressional and shear waves of body-wave seismology, but they differ in that gravitational attraction as well as elasticity provide the restoring forces for the oscillations. The important consequence is that the periods of oscillation of the various modes are directly sensitive to the internal density distribution (see also EARTHQUAKES: *Long-period oscillations of the globe*).

Spheroidal and toroidal modes of free oscillations

There is no simple way to derive the velocity and density distributions directly from the observed periods of free oscillation. Instead, the theory of free oscillations only provides a means of calculating the periods of oscillation if the velocity and density distributions are known ahead of time. Moreover, it is necessary that the velocity distributions satisfy the observed body-wave travel times as well as the free-oscillation periods. Similarly, the density distribution must yield the known values of average density and moment of inertia for the planet.

Effectively what is done is to start with an initial estimate of the density and velocity distributions, calculate the observed properties (periods of oscillations, travel times, and so forth), compare these with the observations, and then refine the density and velocity distributions until the observed properties are satisfactorily reproduced. An elegant mathematical technique, pioneered by the American geophysicists George E. Backus and Freeman Gilbert, makes it possible to "invert" a large number of body-wave, surface-wave, and free-oscillation data in this manner so as to obtain density and velocity profiles that best reproduce the observations. The result is referred to as a seismological Earth model.

Constructing a seismological Earth model

As noted above, determinations of the density, V_P, and V_S profiles make it possible to derive the elastic moduli, K_S and μ, as functions of depth. In addition, the gravitational acceleration at each depth is given by the density profile:

$$g(r) = \frac{4\pi G}{r^2} \int_o^r r'^2 \rho(r')\, dr', \tag{30}$$

where $G = 6.67259 \times 10^{-11}$ $m^3 s^{-2} kg^{-1}$ is the gravitational constant and r is the distance from the Earth's centre (for depth z, $r = 6{,}371 \text{ km} - z$ because the average radius of the planet is 6,371 kilometres). Using equation (30), it is possible to calculate the change in pressure (dP) with a change in depth (dz) according to

$$dP = \rho g dz. \tag{31}$$

Clearly the pressure increases continuously with increasing depth (both ρ and g are positive), but, because ρ and g each depend on depth, the rate at which pressure increases varies with depth.

A technical point worth making is that equation (31) is valid only if the interior is hydrostatic (*i.e.*, shear stresses must be negligibly small, as in a fluid at rest). For this to hold, pressure must vary only with depth; pressure cannot vary laterally at a given depth. In fact, the interior is sufficiently hot that the rock is weak, and only small shear stresses can be sustained over geologic time periods. Maximum shear stresses in the mantle are estimated to be about 0.1 gigapascal, a value that is negligible in comparison with the pressures of the mantle and core. This justifies the use of equation (31) for determining the pressures within the Earth from the seismologically derived density profile. In a different context, however, the small deviations from hydrostaticity are extremely important, as it is exactly these deviations (or the corresponding shear stresses) that drive convection in the mantle and tectonic motions at the surface.

The first attempts to derive a standard Earth model that would be internationally agreed upon were made in the 1970s. The Preliminary Reference Earth Model (PREM) that emerged was presented in 1981 by the American geophysicists Adam M. Dziewonski and Don L. Anderson. This model is based on thousands of observations: travel times for body waves and periods for free oscillations. In addition, the variation of surface-wave velocities with frequency (or period) is used to constrain the V_P and V_S profiles, especially in the top 300 or so kilometres of the mantle. As the velocities and densities are assumed to vary only with radial distance from the Earth's centre (or equivalently, with depth from the surface) in deriving the Earth model, PREM yields laterally averaged values of properties. The resulting profiles of density and velocities are shown in Figure 28, and Table 2 summarizes a number of the seismologically determined properties as functions of depth.

PREM

The Earth's layered or shell structure, illustrated in Figure 26, is clearly defined by the seismic-wave velocities and density. The dominant boundary of the interior, that between the mantle and core, involves a nearly twofold increase in density, with a sharp drop in wave velocities (V_S going to zero) because of the lack of rigidity in the outer core. Next in order of magnitude are the crust–mantle boundary, the inner core–outer core boundary, and the mid-mantle discontinuities at 670 and 400 kilometres depth.

The lateral averaging of properties is more problematic at shallower depths, particularly for the crust and upper mantle, than deeper in the Earth. First, the distinction between oceans and continents (in roughly a proportion of 2 to 1 by area) is averaged out. Second, the variations in crustal thickness are also averaged. The crust–mantle boundary is conventionally taken at the Mohorovičić discontinuity, where the compressional-wave velocity increases sharply to values of eight kilometres per second or higher. The depth to the Mohorovičić discontinuity (*i.e.*, the thickness of the crust) is systematically greater under the continents than under the oceans: on average 35 kilometres and 6–7 kilometres, respectively. In fact, the crust can be as thick as 50–70 kilometres in some continental locations, while being less than a few kilometres thick in others (for example, the Himalayas as compared with parts of northeast Africa).

Thickness of the Earth's crust

The techniques of deep seismic sounding, carried out mainly in the United States and western Europe, have yielded high-resolution velocity profiles for the crust and uppermost mantle since the late 1960s. These studies, complemented by detailed waveform analyses and seismic tomography, prove conclusively that the crust and upper mantle can vary significantly in properties over horizontal distances of tens to hundreds of kilometres. Typically, the lateral variations in velocity are correlated with the tectonic activity that is observed at the surface; for example, young ocean basins (less than 20,000,000 years old), old ocean basins (more than 60,000,000 years old), cratons (continental crust older than 1,000,000,000 to 2,000,000,000 years in age), and tectonically active continental regions commonly exhibit different velocity profiles to depths of 200 kilometres or more.

Among the most notable examples of laterally variable structure is the presence of a low-velocity zone that is typically observed between about 60 and 150 kilometres beneath tectonically active regions, such as the western United States. Here V_P and V_S actually decrease over a limited depth range before resuming their usual increase with depth. The relatively high-velocity mantle overlying the low-velocity zone (wherever present) is referred to as the "lid." The lithosphere, which comprises the tectonic plates at the surface, is sometimes associated with the high-velocity crust and lid.

The sharp discontinuities in the seismic-wave velocities at depths of 400 and 670 kilometres separate the upper and lower mantle. The transition zone between the two discontinuities is a region of anomalously rapid increase in wave velocities with depth, as compared with the rest of the mantle.

Except for its top 100 kilometres and bottom 150–200 kilometres, the lower mantle is notable for its smooth,

continuous increase in velocities with depth. This is in contrast with the much larger velocity variations observed with depth in the overlying upper mantle and transition zone. Also, lateral variations in velocity averaged over 1,000-kilometre distances drop from about 4 to 8 percent in the upper mantle to about 1 to 2 percent in the lower mantle. These observations suggest that the lower mantle is a homogeneous region consisting essentially of a single rock type, whereas the upper mantle is not so uniform in bulk composition or mineral content.

Studies by Bullen and Birch in the 1940s and 1950s yielded a quantitative index of the homogeneity of a region. The local increase in density with increasing pressure is given by the incompressibility, K_S, which can be derived from the measured wave velocities (see equations 27, 28, and 29). As noted above, this was the basis prior to the study of free oscillations for calculating the density profile through regions that are assumed to be uniform in composition. With the free-oscillation measurements, however, the density profile with depth is obtained independently from the wave velocities. Expressing this density variation as a function of pressure rather than depth (see equations 30 and 31), an effective incompressibility can be defined by the relative change in density actually observed inside the Earth with changing pressure:

$$K_E = \frac{dP}{d\rho/\rho}. \tag{32}$$

The inhomogeneity parameter is then defined as the ratio of the density change predicted from the wave velocities to that actually observed from the free oscillations:

$$\eta = \frac{K_S}{K_E} = \frac{\varphi}{\rho g}\left(\frac{d\rho}{dz}\right). \tag{33}$$

For a homogeneous region, η is equal to one. If, however, composition or mineral phases vary with depth, then the density variation is not just determined by elastic compression and η is no longer close to one. Good examples of inhomogeneous regions include the upper mantle and transition zone (Table 2). In contrast, the lower mantle appears to be homogeneous in accord with the inference based on the small lateral variations in wave velocities throughout most of this region.

The D″ layer

The bottom 150 to 200 kilometres of the mantle, called the D″ layer, is highly anomalous relative to the rest of the lower mantle. There the average wave velocities appear to change little with depth (Figure 28). Also, detailed regional studies show that the velocity structure within D″ can be locally complex, with high- and low-velocity zones being present. In addition to the lateral heterogeneity that this implies, evidence that seismic waves can be strongly scattered in the D″ region suggests that large variations in velocity can occur over short distances. Possible explanations of these anomalies include the core–mantle boundary being corrugated with a few kilometres' "topography" (rather than being smooth), and the D″ layer being a region of chemical reaction or physical intermixing between the core and mantle.

Table 2: Structure and Properties of the Earth's Interior

region	depth (z, km)	pressure (P, GPa)	density (ρ, Mg/m³)	gravity acceleration (g, m/s²)	incompressibility (K_S, GPa)	rigidity* (μ, GPa)	quality factor* (Q, in shear)	inhomogeneity parameter (η)
Crust	3	0.03	2.60	9.82	52	27	600	0
	15	0.3	2.90	9.83	25	44	600	0
Upper mantle (10.3%†)								
Lid	24	0.6	3.38	9.84	132	68	600	−0.13
Low-velocity zone	80	2.5	3.37	9.80	130	67	80	−0.13
	220	7.1	3.36	9.90	127	65	80	−0.12
	220	7.1	3.44	9.90	153	76	140	0.78
	400	13.4	3.54	9.97	174	81	140	0.83
400-km discontinuity								
Transition zone (7.5%†)	400	13.4	3.72	9.97	190	91	140	1.73
	500	17.1	3.85	9.99	218	105	140	1.86
	600	21.0	3.98	10.00	249	121	140	0.37
	670	23.8	3.99	10.01	256	124	140	0.37
670-km discontinuity								
Lower mantle (49.2%†)	670	23.8	4.38	10.01	300	155	310	0.98
	770	28.3	4.44	10.00	313	173	310	0.97
	1,000	38.8	4.58	9.97	352	186	310	0.98
	1,250	50.3	4.72	9.94	393	201	310	0.99
	1,500	62.3	4.86	9.93	434	215	310	0.99
	1,750	74.4	4.99	9.95	473	229	310	0.99
	2,000	87.1	5.12	9.99	514	244	310	1.00
	2,250	99.9	5.25	10.08	554	258	310	1.00
	2,500	113.6	5.37	10.25	598	273	310	1.00
D″ layer	2,750	127.6	5.50	10.48	642	287	310	1.01
	2,890	135.8	5.57	10.68	656	294	310	0.99
Outer core (30.8%†)	2,890	135.8	9.90	10.68	644	0	0	0.98
	3,000	147.5	10.08	10.43	686	0	0	0.99
	3,250	172.4	10.44	9.85	777	0	0	1.00
	3,500	200.0	10.77	9.25	862	0	0	1.00
	3,750	222.8	11.06	8.55	948	0	0	1.00
	4,000	246.0	11.32	7.84	1,026	0	0	1.00
	4,250	268.3	11.55	7.13	1,096	0	0	1.00
	4,500	287.3	11.75	6.37	1,160	0	0	1.00
	4,750	304.7	11.93	5.49	1,220	0	0	1.00
	5,000	320.7	12.09	4.84	1,275	0	0	1.01
	5,150	328.9	12.17	4.40	1,305	0	0	1.03
Inner core (1.7%†)	5,150	328.9	12.76	4.40	1,343	157	85	1.00
	5,250	334.4	12.82	3.98	1,356	160	85	0.99
	5,500	345.9	12.92	3.14	1,384	166	85	0.99
	5,750	354.7	13.00	2.26	1,404	171	85	0.99
	6,000	360.4	13.06	1.39	1,417	174	85	0.99
	6,250	363.5	13.09	0.44	1,425	176	85	0.99
	6,371	363.9	13.09	0	1,425	176	85	0.99

*Values at a reference frequency of one hertz (one second period of oscillation) according to the Preliminary Reference Earth Model.
†Mass fraction of Earth.

Characteristics and structure of the Earth's core

Like most of the lower mantle (excluding the D'' layer), the core appears to be a relatively uniform region. It is perhaps for this reason, as well as the large contrast in properties across the core–mantle boundary, that the structure of the core had already been well determined by the 1940s. The presence of the core was inferred by the British scientist R.D. Oldham in 1906, and the depth to the core–mantle boundary (2,900 kilometres) was determined by 1914 through Gutenberg's work. In 1936 the Danish seismologist Inge Lehmann showed that the compressional-wave velocity increases deep inside the core, thus documenting the boundary between the central inner core and the surrounding outer core. Around 1970 the analysis of free oscillations provided independent and definitive confirmation for the presence and finite rigidity of the inner core, surrounded by a fluid outer core. In addition, the density profile derived from the free-oscillation data yields $\eta = 1$ for most of the core (Table 2). The homogeneity that this implies for the outer core is consistent with what would be expected for a fluid region that is being vigorously mixed by convection.

Electrical conductivity. The electrical conductivity at the Earth's surface is extremely variable and is largely dominated by the presence of water. Laboratory measurements, for example, show that the conductivity of seawater is close to four siemens per metre (S/m; or a resistivity of 0.23 ohm-metre), whereas dry rock exhibits conductivities in the range 10^{-11} to 10^{-9} siemens per metre. When water is present in rock, especially in interconnected pore spaces between the grains, the conductivity can range between 10^{-4} and one siemens per metre: approximately a billionfold enhancement over dry rock. Consequently, even though rock itself is electrically insulating, the conductivity of the crust is dependent on the occurrence of water within the rock matrix. In addition, high values of 10^5 siemens per metre or more can be found very locally because of the presence of metallic ores at the surface. Thus, averaging over this enormous spread of values, the near-surface conductivity of the Earth is in the range of 0.1 to one siemens per metre.

The effect of temperature on mantle conductivity

Mantle rock has a finite conductivity because many of the constituent minerals are semiconducting (*i.e.*, their conductivity is intermediate between values typical of insulators and metals, and it increases rapidly with increasing temperature). Water is thought to play a much less important role in the mantle because little is present, as compared with the crust. If temperatures become high enough to partially melt the rock, however, the melt can play the same role as water does at shallower depths in enhancing the conductivity. Therefore, mantle conductivity is expected to be sensitively dependent on temperature at depth.

The electrical conductivity of the mantle cannot be obtained by direct observation. Rather, it is derived by an indirect technique that is based on analyzing temporal variations in the Earth's magnetic field. Specifically, a mathematical description of the geomagnetic field, termed a spherical-harmonic expansion, makes it possible to separately identify that part of the field produced inside the Earth from the part produced externally. In 1838 the German mathematician Carl Friedrich Gauss used this approach to show that more than 95 percent of the field observed at the surface is internally produced. The approach, then, is to analyze the fluctuations of the internal field with time.

Because of the finite conductivity of the mantle, fluctuations in the external magnetic field induce electrical currents in the mantle in accordance with Maxwell's laws of electromagnetism. The electric currents, which are dependent on the conductivity of the mantle, in turn create a fluctuating magnetic field that is now of internal origin. Thus, the value of electrical conductivity in the mantle is derived by comparing the timing of external magnetic-field fluctuations with the induced internal-field fluctuations. In general, fluctuations of longer period are sensitive to conductivity at greater depths.

With this technique, the conductivity of the uppermost mantle has been found to be near 10^{-3} siemens per metre. Beneath continents the conductivity is roughly constant throughout the upper mantle, but beneath young ocean basins there appears to be a region of high conductivity (about 0.1 siemens per metre) between depths of 100 and 200 kilometres. Through the transition zone and the top of the lower mantle the conductivity increases, reaching values of one siemens per metre or more at a depth of 1,000 kilometres. Beyond this depth, conductivity values are not resolved because it is no longer possible to uniquely identify the long-period fluctuations of the internal field due to externally induced currents.

A similar approach can be applied to the lowermost mantle because the internal field generated in the core also fluctuates with time. Depending on the conductivity of the mantle, these fluctuations can be transmitted to the surface: the smaller the conductivity near the bottom of the mantle, the shorter are the periods of fluctuations that can reach the surface. Thus, the observation that internal magnetic-field fluctuations originating from the core seem to occur over periods longer than $3\frac{1}{2}$ to four years can be used to constrain the conductivity to be in the range of 100 to 1,000 siemens per metre in the lowermost mantle.

High conductivity of the core

No direct measurements exist for the conductivity of the core, but this region must necessarily be highly conducting (metallic) in order to produce the geomagnetic field by a dynamo process (see above *The geomagnetic dynamo*). As will be discussed later, the core is thought to consist of an iron-rich alloy. The electrical conductivities of a number of plausible alloy compositions (*e.g.*, iron-sulfur, iron-silicon, and iron-oxygen compounds) have been experimentally measured at the pressures and temperatures existing in the core. The results uniformly point to a value about 10^6 to 10^7 siemens per metre being appropriate for the conductivity of the outer core. A summary of the estimated profile of electrical conductivity through the crust, mantle, and top of the core is included in Figure 29.

Temperature distribution. The temperature distribution through the Earth is mainly constrained by the seismological observations of which regions are solid and which are liquid—*i.e.*, which regions transmit shear waves and

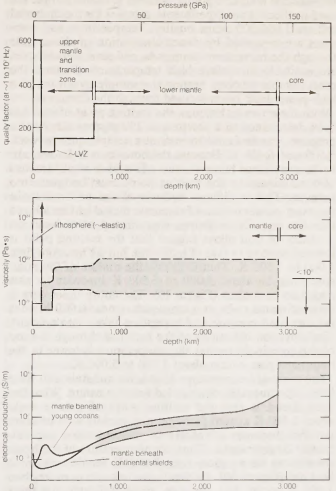


Figure 29: Profiles of the quality factor (Q; see Table 2), viscosity, and electrical conductivity as functions of depth. The quality factor is determined for shear waves at frequencies of one to 100 hertz (periods of one to 0.01 second).

which do not (Figures 26 and 28). Thus, the average temperature in the solid crust and mantle must be lower than the melting point of rocks that make up these regions. Similarly, the average temperature throughout the liquid outer core must be above the melting point of its constituent iron alloy. Presumably, the boundary between the solid inner core and liquid outer core is right at the melting temperature of this alloy.

As the core is considered to be made of nearly pure iron (see below *Zonal variations in chemical and mineralogical composition*), the temperature in the central portion of the Earth can be determined by studying the melting point of iron at high pressures. Even at the top of the core the conditions are fairly extreme, with pressures exceeding 130 gigapascals (Table 2). Nevertheless, two types of experiments have been carried out to determine melting points at these conditions: shock-wave experiments and diamond-cell experiments.

Shock-wave experiments

Shock-wave experiments involve hitting a rock sample with a projectile that is traveling at velocities of several kilometres per second. Upon impact, there is a brief period in which very high pressures and temperatures are achieved in the sample; typically, this lasts less than 0.000001 second before the sample is destroyed by the impact. In this brief period it is possible to measure the temperatures and pressures achieved, as well as the sound velocity through the sample. Thus when high enough temperatures are generated by this shock-compression to melt the sample, the sound velocity is seen to drop. This occurs because of the loss of rigidity on melting, and it is completely analogous to the drop in seismic-wave velocities observed from the inner core to the outer core (Figure 28).

Diamond-cell experiments

In diamond-cell experiments high pressures are achieved by pinching a sample between the points of two diamonds. Diamond is so strong (and hard) that pressures of several hundred gigapascals can be reached in this way. The advantage over shock-wave experiments is that the sample can be observed for hours or days at a time. The disadvantage, however, is that only small samples can be fitted between the diamond points: whereas grams of sample are used in shock-wave experiments, samples for the diamond cell are 1,000,000 times smaller. Temperatures and pressures are measured by direct observation of the sample through the transparent diamonds, and several techniques are available to achieve high temperatures. Perhaps the most notable involves heating the sample by means of a high-power laser beam focused through the diamonds.

With these two techniques, the melting point of iron has been determined to a pressure of 250 gigapascals. At this pressure, iron is found to melt at a temperature between 5,600 and 6,400 K. Because the outer core is known not to be pure iron, however, the effect of a small amount of contaminant on the melting point must be taken into account. This is difficult because of the uncertainty in the exact concentration of elements that might be alloyed into the core. Nevertheless, experiments carried out on a variety of iron alloys suggest that the melting point is lowered relative to that of pure iron, but by no more than 1,000 to 1,500 K. That is to say, the middle of the outer core must be above 4,000 to 5,500 K in order for this region to be molten (Figure 30). The entire inner core is then estimated to be at a temperature near 6,000 K once the increased pressure at this depth is taken into account. Similarly, considerations of the heat lost through the top of the core indicate that the temperature is lower in the outermost core—from about 3,700 to 5,000 K.

The melting temperatures of mantle minerals also have been experimentally determined under pressure. What has been found is that the melting point of mantle rock, about 1,500 K at zero pressure, rises rapidly with pressure but then becomes insensitive to pressure above approximately 20 gigapascals. Thus, the temperature of the lower mantle can be no more than 3,200 K for this region to be solid (Figure 30). Evidently the mantle is more than 1,000 K cooler than the core, which indicates that there must be a sharp increase in temperature near the core–mantle boundary.

Temperature difference between the mantle and core

Similarly, a rapid increase in temperature is known to occur through the crust and uppermost mantle. Very near the surface it is possible to directly measure the increase in temperature with depth; globally, the average increase is between about 10 and 30 K per kilometre (30 to 90° F per mile). Using this value and knowing from laboratory measurements that the thermal conductivity of rock is close to three watts per kelvin metre, the average rate of heat loss from the Earth's interior is calculated to be about 0.07 watt per square metre.

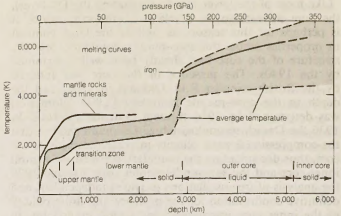


Figure 30: Estimated profile of average temperature at each depth, with the width of the shaded band illustrating the uncertainties. The curves in boldface show the melting temperatures that have been experimentally determined for minerals and rocks occurring in the mantle and for iron at high pressures.

That the temperature in the deep crust and uppermost mantle is high is also documented by geologic observations. The detailed composition of the minerals making up a rock, for example, is usually sensitive to the temperature and pressure at which the rock formed. Rock fragments brought up volcanically from the upper mantle (xenoliths) typically contain olivine, pyroxene, and garnet minerals (see below). It has been found by laboratory experiments that if these minerals are heated together, the amount of calcium and magnesium contained in the pyroxene minerals varies systematically with temperature. In addition, increasing pressure results in increasing amounts of aluminum being present in the pyroxenes. Using these experimental results, the compositions of minerals in naturally occurring xenoliths can be interpreted in terms of their temperatures and pressures of formation. Recalling that pressure increases with increasing depth (see Table 2; Figure 28), one can find that temperatures between 1,000 and 1,800 K occur at depths of 100 to 250 kilometres.

The increase in temperature with depth is so rapid that if it continued unabated the melting curve of the mantle would be crossed by about 400 kilometres depth. Instead, as shown in Figure 30, the average temperature approaches the melting curve to about 150 kilometres beneath the surface and then changes little with depth. This sharp bend in the temperature profile is what keeps the mantle from being molten. Evidence for the bend comes from studies of mineral reactions at high pressures, which indicate that temperatures in the transition zone are roughly 1,500 to 2,000 K (see below).

The overall profile of average temperature with depth is characterized by "kinks"—regions of rapid increase with depth between which the temperature profile is relatively flat. The significance of these kinks became evident during the late 1960s when it was recognized that a vigorously convecting fluid has precisely such a temperature profile. As noted before, the mantle is solid but behaves like a fluid (*i.e.*, it convects vigorously) over geologic time periods.

In the interior of a fluid, corresponding to the interior of either the mantle or core, hot material rises and cold material sinks, thereby transporting (convecting) heat toward the surface very efficiently. Through this interior region of vertical motion, temperature changes little with depth. In contrast, fluid at the surface can move only horizontally; it is the near-surface horizontal movement of the mantle that produces plate tectonics (see below *The cause of plate motions*). In this region heat is transported horizontally, and the only way that the heat can escape outward is to be vertically conducted through the rock. As rocks

Heat flow in the mantle

are good thermal insulators, conduction is very inefficient (compared with convective transport of heat) and a large temperature gradient results with depth. One way to think of this is that though the surface of the Earth is cool, rock provides such good thermal insulation that the mantle close by can be very hot. This conductive region in which the temperature increases rapidly with depth is called a thermal boundary layer. Such a layer apparently occurs at the core–mantle boundary as well as at the surface.

Temperature variations near the Earth's surface

It is important to note that so far only the average temperature at a given depth has been considered. In fact, temperature varies considerably in the horizontal direction, as well as vertically, especially near the surface. Temperature differences as large as 1,200 K, for example, are present between volcanic areas in which lava (molten rock) is present and the rest of the crust. Globally, the most important volcanic structure is the mid-oceanic ridge system, at which new oceanic crust is produced. The rock underlying the oceanic ridges forms the hot, upwelling part of the mantle-convection pattern. In contrast, the lithosphere that sinks into the mantle at oceanic trenches (the subduction zones) forms the cold, downwelling part of the convection. This subducted lithosphere becomes warmed up as it sinks into the mantle. Thus, the variation in temperature at a given depth is greater than 1,000 K in the uppermost mantle, but it decreases with depth, becoming three or four times smaller in the lower mantle. In the core, horizontal temperature variations are expected to be 10 to 100 times smaller still, because the outer core is thought to have a low viscosity (*i.e.*, be extremely fluid).

Rheological properties. Rheology refers to the ease or difficulty with which a material deforms permanently rather than elastically. For a fluid this is characterized by viscosity, the ratio of the shear stress required to produce the deformation divided by the rate of the deformation. Solids can be characterized by a yield strength, given by the minimum stress required to produce permanent deformation (*e.g.*, fracturing). On geologic time scales, however, solid rocks at high temperature deform even under low shear stresses. Therefore, the rock can be thought of as having an effective (fluid-like) viscosity defined again as the ratio of the shear stress to the deformation rate.

Seismic anelasticity

Two effects by which the rheology of the interior is evaluated are seismic anelasticity and postglacial rebound. The first is based on the fact that minerals do not behave perfectly elastically in response to seismic waves but are slightly "mushy" under the deformations caused by the waves. As a result, seismic waves are attenuated, becoming smaller in amplitude as they travel through the interior. Similarly, free oscillations become damped with time. A quantitative measure of this anelasticity is the quality factor

$$Q = \frac{\pi}{\delta A/A}, \tag{34}$$

Here, $\delta A/A$ is the relative decrease in amplitude per oscillation of the seismic wave, and a high value of Q means that there is little attenuation. Also, the deformation energy that is lost to viscous damping is given by $2\pi/Q$ per cycle of oscillation.

Measurements of body-wave amplitudes as functions of distance of travel and observations of the decay rates of free-oscillation modes are used to derive a profile of Q with depth. The quality factor for shear deformation, which is inversely proportional to the amount of shear-wave attenuation, is shown in Figure 29 and included in Table 2. Most notable are the relatively high value near the surface (in the crust and mantle lid), the low value in the low-velocity zone, and the increasing values with further depth into the mantle. Shear Q is zero in the outer core, and it appears to be relatively small in the inner core (Table 2). By the definition of the quality factor, therefore, the most nearly elastic portion of the Earth is the high-Q zone near the surface.

Postglacial rebound

Postglacial rebound involves studying the deformation caused by Pleistocene (Ice Age) glaciers that covered much of North America and Scandinavia. The weight of the ice in these glaciers depressed the surface of the Earth by many hundreds to thousands of metres. As the load of ice has been removed by melting over the past 100,000 years, mantle rock is flowing back under the depression. Hence, the surface is moving upward and the depression is vanishing at a rate that is determined by how slowly the mantle flows—*i.e.*, how viscous the mantle is.

The rate of postglacial uplift is determined by measuring the ages of ancient beaches and the amount by which these beaches have been elevated above sea level. Typical rates are in the range of a few metres per 1,000 years. From such measurements at several locations, the viscosity profile through the upper mantle can be derived. What is found is that there is a low-viscosity layer underlying a surface layer that behaves elastically (Figure 29). The lithosphere is defined as being this top elastic layer, and it corresponds closely to the high-Q crust and lid that are observed seismologically. At depths beyond the transition zone, the viscosity profile is poorly resolved because motions due to postglacial uplift become vanishingly small.

As an illustration of how viscosity estimates can be used, consider the convective deformation of the mantle that is believed to cause plate tectonics. Taking plate velocities to be centimetres per year (10^{-2} metre in 3×10^7 seconds) and assuming that the deformation is across the approximately 3,000-kilometre thickness of the mantle, this yields a deformation rate of $(10^{-2}\ \text{m}/3 \times 10^7\ \text{s}/3 \times 10^6\ \text{m} =) 10^{-15}\ \text{s}^{-1}$. Thus, with a mantle viscosity of about 10^{21} Pa · s (Figure 29), shear stresses of $10^{21}/10^{15} = 10^6$ Pa (pascals; or about 10 atmospheres) are sufficient to drive the observed tectonic motions. Compared with the mantle pressures listed in Table 2, the shear stresses are negligibly small. As previously indicated, the interior is almost perfectly hydrostatic, even though it is the deviations from hydrostaticity that cause tectonic motions.

The most interesting correlation among the profiles shown in Figures 28 and 29 is the occurrence of the low-Q, low-seismic-velocity, low-viscosity, and high-electrical-conductivity layers at about the same depth in the upper mantle. Apparently all of these measurements indicate that the rock is relatively soft or fluid at this depth. Referring to the temperature profile (Figure 30), this makes sense because it is precisely in this depth range that the rocks are brought most nearly to the melting point. Close to the melting temperature the rock would be expected to be softened and, especially if a small amount of partial melting occurs, the high electrical conductivity beneath young oceanic crust is similarly explained. Thus, there is a major rheological contrast between this layer, the so-called asthenosphere, and the relatively stiff lithosphere above it.

Presumed low viscosity of the core

The viscosity of the core is not well known but is thought to be low. By the mid-1980s, astronomical observations of the wobbling in the Earth's rotation could be interpreted to give an upper limit of about 10^3 pascal · seconds for the outer core. Measurements of the viscosity of molten iron extrapolated to high pressures, however, indicate that the outer core is likely to be at least 1,000 times less viscous.

ZONAL VARIATIONS IN CHEMICAL AND MINERALOGICAL COMPOSITION

Most of the Earth's surface is covered by oceanic crust. The composition of this crust is documented on the basis of rocks sampled by deep-sea dredging and by studies of ophiolites, which are slivers of oceanic crust and uppermost mantle that have been tectonically emplaced on land. The oceanic crust is found to be relatively uniform in composition (Table 3). It consists almost entirely of basalt (or the fully crystallized equivalent, gabbro), a rock containing pyroxene, plagioclase, olivine, and iron-oxide minerals, along with glass (see Table 4). The basalt is erupted at mid-oceanic ridges, solidifying to form new crust that is transported away from the ridge in the plate-tectonic cycle. In actuality, much of the oceanic basalt that is found has reacted with seawater after its formation and contains hydrous minerals such as serpentine $([Mg,Fe]_3Si_2O_5[OH]_4)$.

The continental crust is considerably more heterogeneous than oceanic crust and differs systematically from the latter in composition. Attempts to average the compositions of all continental rocks observed at the surface

Table 3: Estimates of Average Composition (weight percent)

oceanic crust*		continental crust*		mantle*†		core
SiO_2	49.5	SiO_2	57.3	SiO_2	45.0	Fe ~ 90
TiO_2	1.5	TiO_2	0.9	TiO_2	0.2	S, O ~ 5–10
Al_2O_3	16.0	Al_2O_3	15.9	Al_2O_3	4.5	
Cr_2O_3	—	Cr_2O_3	—	Cr_2O_3	0.4	
FeO	10.5	FeO	9.1	FeO	7.6	
MnO	—	MnO	—	MnO	0.1	Other plausible constituents: Ni, Si, H
NiO	—	NiO	—	NiO	0.2	
MgO	7.7	MgO	5.3	MgO	38.4	
CaO	11.3	CaO	7.4	CaO	3.3	
Na_2O	2.8	Na_2O	3.1	Na_2O	0.4	
K_2O	0.2	K_2O	1.1	K_2O	0.01	

*From S.R. Taylor and S.M. McLennan. *The Continental Crust: Its Composition and Evolution* (1985). †After Basaltic Volcanism Studies Project, *Basaltic Volcanism on the Terrestrial Planets* (1981).

Difficulties in determining the average composition of the continental crust

or of sediments formed by erosion of large continental regions do not yield good estimates of the average crustal composition. This is because rocks formed in the deep continental crust are less often exposed at the surface than are those from shallower depths. Furthermore, the former appear to be systematically more depleted in silicon and enriched in calcium and aluminum than are the latter. Correcting for these differences leads to the estimate of average composition listed in Table 3. The composition of the continental crust corresponds roughly to that of andesite, which is considerably richer in silicon than the basalt constituting oceanic crust. Therefore, the deep continental crust is roughly intermediate in composition between the shallow continental crust and oceanic crust. Among the dominant minerals making up the continents are feldspars (mainly plagioclase), quartz, micas, and amphibole ($[Mg,Fe,Ca,Na]_{2-3}[Mg,Fe,Al]_5[Si,Al]_8O_{22}[OH]_2$).

The composition of the upper mantle is less easily determined than that of the crust. Samples of mantle rock come mainly from two sources: ophiolites and xenoliths. The latter, in particular, include rocks that have been brought from great depth by explosive eruptions. Most notably, the occurrence of diamonds with some xenoliths requires a source beyond 150 kilometres depth. As diamond is a high-pressure form of carbon, it converts rapidly to graphite when heated at a pressure less than five gigapascals, corresponding to this depth. Xenoliths range in composition from dunite, a rock made almost entirely of olivine, to eclogite. Eclogite is a variety of rock that is formed when basalt is taken to pressures exceeding 1.5 gigapascals; at such pressures the aluminum-bearing minerals, plagioclase and pyroxene in basalt, react to form new pyroxene and garnet in the eclogite. Common in xenoliths is peridotite, which consists mainly of olivine, pyroxene, and lesser amounts of plagioclase or garnet. As with basalt and eclogite, plagioclase is present if the peridotite is crystallized at low pressures and garnet is present if crystallized at higher pressures.

Peridotite as the dominant rock of the upper mantle

Based on the observed samples, peridotite with about 50 percent olivine, 30 percent pyroxene, and 15 percent garnet is considered to be the dominant rock of the upper mantle. This conclusion is supported by melting studies that demonstrate that a basaltic liquid is produced if peridotite is heated enough to melt partially. It is by partial melting of the upper mantle beneath mid-oceanic ridges that the basalt composing the oceanic crust is produced. The effect of this partial melting is to slightly deplete the original peridotite in aluminum and iron—elements that preferentially are concentrated into the basaltic liquid. Thus, the more aluminum- and iron-rich xenoliths are sometimes referred to as "fertile," meaning that they have not yet been significantly melted and could produce basaltic liquid upon partial melting.

Numerous experiments have been carried out at high pressures and temperatures to document the changes in mineral content and physical properties that occur in peridotite as it is taken to greater depths in the mantle. Among the most important changes are transformations of the minerals to denser crystal structures under pressure. In the 1960s and '70s, for example, studies by the Australian geochemist Alfred E. Ringwood and the Japanese geophysicist S.I. Akimoto showed that the pyroxene in peridotite transforms to the structure of garnet (the mineral is named majorite) and that the olivine transforms to phases with spinel-like crystalline structures (β-phase and γ-spinel, which have the mineral names wadsleyite and ringwoodite, respectively; see also MINERALS AND ROCKS: *Major rock-forming minerals: Olivines*).

The findings of Ringwood and Akimoto

One of the remarkable findings was that the transformation of olivine to its high-pressure (spinel) forms involves a jump in the elastic-wave velocities and density. Also, the transformation occurs suddenly at about the pressure of the 400-kilometre seismological discontinuity (Table 2). The boundary between the upper mantle and transition zone therefore takes on a natural explanation as being the depth at which the crystal structure of olivine that dominates the upper mantle transforms under pressure. Furthermore, the pressure at which this transformation occurs depends somewhat on temperature, an effect that has been experimentally calibrated. Thus, olivine transforms at exactly the pressure of the 400-kilometre discontinuity at a temperature of 1,700 K—a result that directly constrains the temperature in the transition zone, as noted above. In contrast with the olivine-spinel transition, the breakdown of pyroxene to form majorite occurs over an extended pressure range (rather than suddenly), and this can explain the anomalous increase in wave velocities through the transition zone (Figure 28).

In collaboration with Ringwood at the Australian National University, Lin-Gun Liu demonstrated in 1975 that

Table 4: Properties of Crust, Mantle, and Core Minerals*

mineral phase	chemical formula	pressure-range of existence (P, GPa)	density (ρ, Mg/m^3)	incompressibility (K_S, GPa)	rigidity (μ, GPa)	thermal diffusivity†‡ κ, 10^{-5} m^2s^{-1})	thermal expansion coefficient† (α, 10^{-5} K^{-1})
Quartz	SiO_2	0 to 2	2.65	38	44	—	−0.3
Feldspar (plagioclase)	$[NaSi,CaAl]AlSi_2O_8$	0 to <15	2.62–2.76	55–92	29–41	—	2.5
Mica (muscovite)	$KAl_2(AlSi_3O_{10})(OH)_2$	0 to <10	2.79	57	34	—	—
Coesite	SiO_2	2 to 8	2.92	98		3	1.4
Olivine	$(Mg,Fe)_2SiO_4$	0 to 13	3.22	129	81	0.7	4.0
Pyroxene	$(Mg,Fe,Ca)SiO_3$	0 to 13	3.21	108	76	1–2	4.2
Garnet	$(Mg,Fe,Ca)_3Al_2Si_3O_{12}$	0 to 30	3.56	174	89	1–2	2.9
Majorite	$(Mg,Fe,Ca)_3[(Mg,Fe)Si]Si_3O_{12}$		3.51	220			
β-phase	$(Mg,Fe)_2SiO_4$	13 to 17	3.47	174	114	—§	3.4
γ-spinel	$(Mg,Fe)_2SiO_4$	17 to 16	3.55	184	119	1	2.7
Stishovite	SiO_2	8 to >50	4.29	316	220	2	1.7
Magnesiowüstite	$(Mg,Fe)O$	0 to >100	3.58	163	131	2.6	4.8
Perovskite	$(Mg,Fe,Ca)(Si,Al)O_3$	20 to >120	4.10	262	[155]‖	—§	4.0
ε-iron	Fe	10 to >100	8.35	195	—	—	—
Liquid iron (1,800 K)	Fe	0 to >25	7.02	85	0	—	11.9

*Unless otherwise stated, these are values experimentally determined at zero pressure and room temperature for the magnesium end-member composition. †High-temperature values measured at approximately 1,000 K. ‡Thermal diffusivity is related to thermal conductivity (k) by $k = \kappa\rho C_p$, with C_p (~1 kJ/K · kg for mantle minerals at high temperature) being the specific heat at constant pressure. §Unmeasured value. ‖Unmeasured value estimated on the basis of theoretical calculations.

pyroxene further transforms to a dense phase with the perovskite structure. Liu, one of the pioneers in the application of the laser-heated diamond cell, subsequently showed that olivines, pyroxenes, and garnets of a wide variety of compositions all transform to the perovskite phase (sometimes with a coexisting oxide phase) at pressures above 15 to 25 gigapascals. This turned out to be especially significant when further research conducted in the United States showed that the perovskite phase of silicate minerals is stable to pressures existing at the core–mantle boundary.

Silicate perovskite as the Earth's most abundant mineral

Apparently, peridotite of the upper mantle transforms almost entirely to the perovskite phase at the pressures near the top of the lower mantle, and the resulting perovskite exists throughout the deeper mantle. Because the lower mantle makes up more than 60 percent of the Earth on an atomic basis, silicate perovskite is considered to be by far the single most abundant mineral in the planet.

The crystal structure of perovskite is built up of SiO_6 octahedral units. In contrast, the minerals observed in the crust and upper mantle are characterized by SiO_4 tetrahedral building blocks. The tetrahedrons of the low-pressure structures are usually linked in complex arrangements, including chains and rings, to form the common minerals. At higher pressures, however, much denser packing of the atoms is achieved by locating six (rather than four) oxygen atoms around each silicon atom and by placing these octahedrons in a tightly packed arrangement. A good example is the mineral stishovite, which is the dense, high-pressure form of quartz (Table 4; stishovite was first synthesized in the laboratory by the Russian physicist Sergey M. Stishov and was subsequently discovered as a shock product in natural meteorite craters). In the mantle, silicon begins to form octahedral units in the garnet-structured majorite that occurs in the transition zone. The dominant change from tetrahedrally to octahedrally bonded silicon, however, occurs at the top of the lower mantle.

The changes in seismological properties with depth through the crust and mantle can be well understood either in terms of pressure-induced structural transformations, such as those just described, or in terms of relatively small changes in composition, such as the difference between the crust and mantle (see Table 3). In his studies of the Earth's interior, however, Birch already recognized that the change across the mantle–core boundary must be much more profound. Considering the seismological data alone, this boundary stands out as involving a large decrease in the seismic parameter, from $\varphi = 118$ km^2/sec^2 at the bottom of the mantle to $\varphi = 65$ km^2/sec^2 at the top of the core. These values are for adjoining regions across the boundary, which are therefore at the same temperature and pressure; also, because φ depends only on incompressibility and density, the lack of rigidity in the outer core does not come into play. For comparison, the seismic parameter changes by less than 15 km^2/sec^2 across the crust–mantle boundary and by only 2 to 4 km^2/sec^2 across the mid-mantle discontinuities and the inner core–outer core boundary.

The radical change in density from mantle to core is apparent in Figure 31. Also included are experimental measurements for a variety of compounds to show that the densities of magnesium, silicon, aluminum, and calcium oxides at elevated pressures and temperatures are in close agreement with the observed lower-mantle densities. This is in accord with the deep mantle consisting of high-pressure oxide or silicate minerals and contrasts sharply with the observed properties of the core.

Aside from having a high density, the generation of the magnetic field also requires that the core, unlike the mantle, be metallic. Birch's conclusion in 1952 was that iron is the most plausible constituent of the core. This was bolstered by subsequent work that has shown that iron is by far the most abundant element in the cosmos with density, elasticity, and electrical properties comparable to those of the core.

As seen in Figure 31, the outer core is about 10 percent less dense than iron at comparable pressures and temperatures. On this basis, the core is inferred to contain alloying elements or contaminants, such as sulfur or oxygen. The reasons for considering these particular elements

Alloying elements in the core

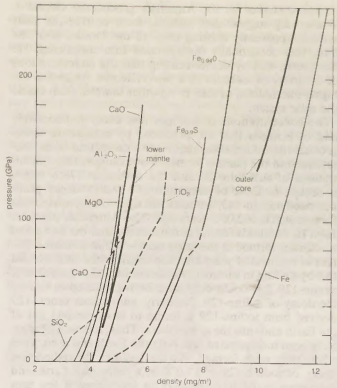


Figure 31: Summary of shock-wave data on the densities of oxides and iron compounds at high pressures and high temperatures. The seismologically derived pressure-density curves for the lower mantle and outer core are included for comparison.

as contaminants are somewhat tentative and are based largely on geochemical considerations, such as their cosmic abundances. Both iron oxide and sulfide are metallic at the conditions of the core, and both exhibit somewhat lower densities. By comparing densities of the iron compounds at elevated pressures and temperatures, it is found that the seismologically observed profiles of velocities and densities through the core can be matched by an alloy composition of about 90 percent iron by weight, with the remainder being sulfur, oxygen, and a variety of possible minor components (Table 3).

DEVELOPMENT OF THE EARTH'S STRUCTURE AND COMPOSITION

The origin of the Earth in its present form has been the subject of intellectual interest for centuries, but the decades since 1950 have seen major advances both in concepts and in measurements pertaining to this topic. The quantitative analysis by geochemists of the isotopes in meteorites and, in particular, the study of lunar rocks obtained during the U.S. Apollo program have produced some of the major contributions. In addition, geochemical research on terrestrial samples, combined with the new understanding of internal processes brought on by the recognition of plate tectonics, have significantly elucidated the ways in which the Earth has evolved.

The significance of nucleosynthesis

The astrophysical concept of nucleosynthesis remains the starting point in tracing planetary evolution. This includes the nuclear processes by which the light elements—hydrogen through boron—were produced at the birth of the universe, some 10,000,000,000 to 20,000,000,000 years ago. By analogy with what astronomers presently observe to happen, it is thought that the solar system represents preexisting cosmic gas and dust that was collected and compacted to extremely high density by a shock wave emanating from a nearby supernova (violently exploding star). Once sufficiently high pressures and densities were achieved, nucleosynthetic processes could begin within the solar mass to produce the elements that make up the solar system. The birth of the Sun, which makes up more than 99.9 percent of the mass of the entire solar system, is taken to be the time at which the planets started to form approximately 4,600,000,000 years ago.

As the vapour making up the solar nebula expanded and cooled, mineral grains are thought to have condensed and aggregated to form the earliest meteoritic material. In

addition, studies by the American geochemist Gerald J. Wasserburg suggest that material from outside the solar system, apparently existing prior to the formation of the Sun, was occasionally incorporated into meteorites. He has shown this by demonstrating that the concentrations of a variety of isotopes in a few meteorite fragments are highly anomalous, as seen in no other samples from inside the solar system.

Isotopic analyses

The concentrations of isotopes that decay radioactively and of isotopes that are produced by radioactive decay provide the information required to determine when meteorites and the planets formed. For example, the concentrations of rubidium-87 and the strontium-87 to which it decays, or those of samarium-147 and its decay product neodymium-143, indicate that the oldest meteorites formed 4,600,000,000 years ago. The American physicist John H. Reynolds further demonstrated that the Earth and meteorites formed at the same time—within at most a few tens of millions of years after the origin of the Sun. He did this by finding in meteorites anomalously high amounts of xenon-129 that could only have been produced in situ by the decay of iodine-129. Similarly, anomalous xenon-129 derived from iodine-129 is found to have escaped out of the Earth and into the atmosphere. Thus, iodine-129 must have been incorporated into both the Earth and meteorites when they were formed. As the half-life for radioactive decay of iodine-129 is 17,000,000 years, the Earth and meteorites could have formed no more than a few tens of millions of years after nucleosynthesis, when the Sun originated and iodine was formed. Later, essentially no iodine-129 was left to produce the anomalous xenon.

The most abundant elements in the Sun, hydrogen and helium, are severely depleted in the terrestrial planets of the inner solar system but are still abundant constituents of the large, gaseous planets of the outer solar system (*e.g.*, Jupiter and Saturn). It is thought that only in the colder regions of the outer solar system, including the zone beyond Pluto in which comets originated, could these volatile elements be retained during the formation of the planets. Considering only the less volatile (less gaseous) elements, however, it is found that the relative elemental abundances for the Sun, for a class of old and largely unaltered meteorites (called CI chondrites), and for the estimated composition of the Earth are all in close agreement. This

The chondritic model

is the basis for the chondritic model, which holds that the Earth was essentially composed of and has a composition similar to such meteorites. Isotopic analyses carried out in Wasserburg's laboratory during the late 1970s corroborate this idea: rocks derived from regions that are considered to be extremely primitive within the Earth (meaning that the regions have changed little through Earth history) exhibit concentrations of neodymium and samarium isotopes that are virtually identical to values obtained from chondritic meteorites. The reason that neodymium and samarium are considered is that these elements have similar volatilities and are both relatively involatile, making them good geochemical tracers. Thus, it appears that the composition of the Earth is roughly what would be expected given the observed abundances in the Sun, except that the lighter, more volatile elements were not retained very well as the planet formed.

The dust and grains that condensed out of the cooling solar gas aggregated to form larger fragments of rock. Chondritic meteorites are basically just such collections of grains and fragments that have been compacted together to form a larger piece of rock and eventually small planetary bodies. Such bodies are termed planetesimals when they become roughly as large as asteroids (several kilometres to a few hundred kilometres in dimension). The larger they grow, the greater the gravitational attraction that the planetesimals exert and hence the more effectively they sweep up additional particles and rock fragments while circling the Sun.

Both metallic meteorites (those composed largely of iron alloys of nickel and sulfur) and stone meteorites fall on the Earth today, and both types are thought to have been present during the accretion (or aggregation) of the planetesimals that were to form the Earth. That is to say, the Earth seems to have accreted after most, if not all, condensation of solids was complete. Thus, a wide range of minerals was included in the grains, larger fragments, and even planetesimals that were swept up by the growing planet. It seems that such an aggregation of dense metallic fragments and less dense rocky fragments is not very stable. Calculations based on the measured strengths of rocks indicate that the metallic fragments probably sank downward as the Earth grew. Although the planet was relatively cold at this stage (less than 500 K), the rock was weak. This is an important point because it leads to the conclusion that the Earth's core began to form during accretion of the planet and probably before the planet had grown to one-fifth of its final (present) volume.

Shock-heating of the Earth during accretion

During accretion the planet is thought to have been shock-heated by the impacts of meteoritic and planetesimal bodies. For meteorites this heating is concentrated near the surface (where the impacts occur), which is cooled by radiation back into space. The Russian planetary physicist V.S. Safronov has pointed out, however, that the larger planetesimals can penetrate sufficiently deeply upon impact to produce heating well beneath the surface. In addition, the debris formed on impact can blanket the planetary surface, which also helps to retain heat inside the planet. In this way, concludes Safronov, the Earth became hot enough to begin melting after growing to less than 15 percent of its final volume.

Among the planetesimals striking the Earth some 4,600,000,000 years ago, at least one is considered to have been comparable in size to the Earth—a large fraction of the present diameter of the planet. Although the details are not well understood, there is good evidence that the Moon was formed by the impact of such a large planetesimal on the Earth. Specifically, Ringwood has shown that the relative abundances of many trace elements in lunar rocks are close to the values obtained for the Earth's mantle. Unless this is a fortuitous coincidence, it points to the Moon having been derived from the mantle. Impact calculations suggest that a glancing collision of a large asteroid or small planetary body could have been sufficient to excavate the material that would form the Moon from out of the Earth's interior. Again, the evidence for such large collisions points to the Earth having been very effectively heated during accretion.

The conclusion is that many processes occurred almost simultaneously within a matter of tens to hundreds of million years after the Sun was formed. The evidence from xenon isotopes shows that meteorites and the Earth were formed within this time. Apparently the Moon, which is dated as more than 4,000,000,000 years old, was formed from the Earth's mantle in the same time period. Simultaneously, the core was accumulating toward the planetary centre, and may have been completely formed during the growth period of the Earth. In addition to the accretional heating caused by planetesimal impacts, the sinking of metal to form the core released large enough amounts of gravitational energy to heat the entire planet by 1,000 K or more. Thus, once core formation began, the Earth's interior became sufficiently hot to convect. Although it is not known whether or not plate tectonics was active at the surface, it seems quite possible that the underlying mantle convection began even before the planet had grown to its final dimensions.

Once hot, the Earth's interior could begin its chemical evolution. For example, outgassing of a fraction of those volatile elements that had been trapped (in small amounts) within the accreting planet probably formed the earliest atmosphere and oceans at this time. Also, chemical reactions between the mantle and core became possible in the deepest interior. From the perspective of surface geology, however, perhaps the most important event was the formation of crust by partial melting. This chemical separation by partial melting and devolatilization is termed differentiation.

Planetary differentiation

Exactly when and how the continental crust began to grow is uncertain, because the oldest rocks found so far are only 3,900,000,000 years old. These rocks were metamorphosed from preexisting crustal rocks, so it is known that the crust was present earlier in Earth history. In fact, one mineral grain, a zircon from Australia, has been dated

at 4,276,000,000 years in age, but its relation to the formation of continental crust is uncertain. Although direct evidence is not available, modeling of the isotopic compositions of rocks indicates that continental crust formed early. The analysis of neodymium and samarium isotopes and of rubidium and strontium isotopes, for example, suggests that the average age of continental crust is about 2,500,000,000 years. Thus, in all probability, repeated partial melting of the upper mantle formed successively more refined, continent-like crustal rocks starting from before 4,000,000,000 years ago. Over the first 1,000,000,000 years, however, much of the continental crust that was formed appears to have been reincorporated into the mantle: recycling rates of about one-third of the continental crust per 1,000,000,000 years are inferred from the isotopic data. As a result, only a few fragments of crust older than 3,500,000,000 years remain.

The process of partial melting and formation of crust, especially continental crust, leads to a depletion of certain elements (*e.g.*, silicon and aluminum) from the mantle. Relatively primitive and undepleted regions are still present, making up about one-third to one-half of the mantle, according to the isotopic models. The distribution of depleted and undepleted regions, however, is uncertain. Although much (perhaps all) of the upper mantle appears to be depleted, it is not known whether only the upper mantle is depleted or whether depleted rocks also occur in the lower mantle.

Continued chemical and thermal evolution of the Earth's interior

What is recognized is that the Earth is still differentiating, separating into chemically distinct layers or regions. This is most evident in the processes of plate tectonics that involve ongoing extraction of oceanic crust from the upper mantle. Although this crust is eventually subducted, along with overlying continent-derived sediments, remixing of oceanic or continental crustal material back into the mantle appears to be not very efficient. That is, the crustal material is physically returned to the mantle without being completely chemically homogenized. Thus, chemical and thermal evolution of the interior are intimately connected through convection, and the twofold process is still vigorously in progress some 4,600,000,000 years after the formation of the planet. (R.J.)

The major geologic features of the Earth's exterior

DEFORMATION OF THE CRUST

The Earth's crust has been subjected to widespread deformation over geologic time. This deformation results from forces applied to the rocks that make up the crust. Such forces create stresses within the rocks, and these stresses in turn produce strain, which is manifested in the form of folds, faults, and joints.

Force, stress, and strain. *Force.* This is a vector quantity that changes or tends to produce a change in the motion of a body. It is a push or pull and must be specified as to its direction and magnitude. If there is no acceleration of a body, the forces that act on it are balanced. There are four types: compression, tension (or extension), shear, and torsion. During compression, rocks are squeezed together by two equal forces acting toward each other along the same line. During tension, rocks are pulled apart by two forces acting away from each other along the same line. Shear is produced by a force couple; *i.e.*, rocks are subjected to two equal forces acting in opposite directions but not along the same line. Torsion is a twisting action produced by two opposed force couples acting in parallel planes.

Types of forces that deform rocks

Stress. When forces are applied to the external surface of a body, they set up internal stresses within it. Stress can be taken as a measure of the intensity of a force acting within the body. It is properly defined as the force per unit area within a body and is measured in dynes per square centimetre.

Strain. Any deformation caused by stress constitutes strain. It may be an increase or decrease in the length of an object, an alteration of its shape, or a change in its volume. It is generally measured as a percent elongation, percent shortening, angular distortion, or percent change in volume.

Geologic structures. *Folds.* Most folds are produced either by compression or by shearing of the Earth's crust. They generally occur in elongated belts and are produced within mountain systems. Folds in which rock layers are arched upward are called anticlines, and those in which the rock layers are bowed downward are called synclines (Figure 32). Homoclines are structures in which sedimentary rocks dip uniformly in one direction.

Parallel, similar, and flow folds

Folds can be classified on the basis of the form of the deformed rock layers. Parallel (concentric) folds are folds in which the layering in each bed is parallel to the layering in other beds and in which the layering forms concentric circles (Figure 32, left). Similar (slip) folds are those in which the layering in each bed is parallel to the layering in other beds and in which the form is the same in each layer in the fold (Figure 32, centre). Another type of major fold is the flow fold, in which the layering is neither parallel nor of the same form from one layer to another; instead, the layering is disharmonic (Figure 32, right). Flow folds form deep within the Earth's crust under conditions that produce metamorphism in the deformed rocks.

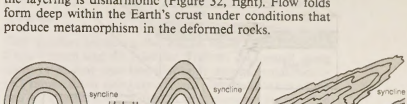


Figure 32: The forms of three types of folds.

Parallel folds usually develop as a result of compressional forces, such as those accompanying the collison of lithospheric plates. They tend to form at the outer margins of a fold belt. The so-called similar folds form due to shear stresses produced by a force couple in which both forces are oriented in a vertical plane. Shear folds often occur beneath major reverse faults or convergent plate margins. In such cases, the force couple that produces shear folds develops as a result of frictional resistance (drag) beneath the fault or plate margin. Flow folds can form either from compressive or shearing forces. As is the case in similar folds, both forces in the force couple that create flow folds are oriented in a vertical plane. Flow folds produced by compressional forces typically develop as a result of plate collisions, while those that form as a result of shearing forces develop beneath convergent plate margins. Flow folds form in the core of mountain belts and are exposed only after extensive erosion of the belts.

Faults. Faults are fractures along which movement has occurred. They are produced by compression, tension, or shearing of the Earth's crust. Faults can be characterized by the direction of movement along the fault plane. A dip-slip fault is a fracture in which the movement is parallel to the dip (direction of tilt) of the fault plane. A strike-slip is one in which the movement is parallel to the strike (direction or trend) of the fault. An oblique-slip fault shows a combination of both dip-slip and strike-slip movements.

Normal and reverse dip-slip faults

Dip-slip faults may be either normal or reverse, depending on the relative movement of the hanging wall and the footwall of the fault. (The hanging wall is the block lying above the fault surface, while the footwall is that which lies beneath the fault surface; see Figure 33.) In a normal fault the hanging wall is displaced downward relative to the footwall, whereas in a reverse fault the hanging wall is displaced upward. A structure bordered by two normal faults dipping toward each other is a graben.

Strike-slip faults may be either right-lateral or left-lateral. In a right-lateral strike-slip fault, the far side of the fault is displaced to the right. The opposite occurs in a left-lateral fault of this type. Most strike-slip faults form because of shearing of the Earth's crust. Such shearing occurs as a result of the development of a force couple in which both forces are oriented in the horizontal plane.

Some faults form as a result of gravitational rather than tectonic stresses. In Wyoming, for example, gravitational force caused large blocks of sedimentary and volcanic

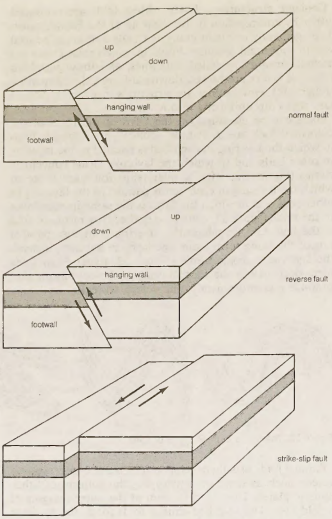


Figure 33: Three basic fault types: (top) normal fault, (middle) reverse fault, and (bottom) strike-slip fault.

rocks to slide tens of kilometres downhill along a fault surface called the Heart Mountain Detachment.

Joints. A joint is a fracture along which there has been no perceptible movement. Joints tend to occur in sets containing numerous subparallel joints, and typically there are two or three dominant joint sets in an area.

Joints have a variety of causes. They may be produced during the cooling of an igneous rock. Such joints are called primary joints, and all others are secondary joints. Cooling of an igneous rock results in its contraction, and this produces tension cracks. Features of this type are referred to as columnar joints if they tend to form long polygonal columns.

Sheeting

Joints also may be produced by the release of pressure on a massive igneous or sedimentary rock, such as granite or thick-bedded sandstone. Such jointing is termed sheeting. It occurs when a rock that has been deeply buried under hydrostatic pressure (pressure equal in all directions) is exposed relatively rapidly by erosion, resulting in a significant reduction in the pressure perpendicular to the erosion surface. Fracturing is caused by residual stresses oriented parallel to the erosion surface.

Another major cause of joints is regional stresses. Both compressive and shear stresses tend to produce joints along which there has been a slight but not perceptible movement. Such joints are called shear joints, and they are incipient faults. Regional tension would produce a vertical joint set oriented at right angles to the direction of extension. Regional torsion generates local tension that may give rise to intersecting joint sets. Experimentally produced torsion fractures closely resemble fractures found in sedimentary rocks in the mid-continental region of North America. Regional torsion may be produced by differential uplift of sedimentary sequences within the mid-continent where domes are uplifted more than the adjacent basins. It also may be caused by Earth tides generated as a result of the gravitational attraction between the solid Earth and both the Moon and Sun. Earth tides lift the outer surface of the planet about 20 centimetres. Such uplifts would cause repeated flexing of the Earth's crust and could induce jointing in sedimentary layers.

PHYSIOGRAPHIC EXPRESSIONS OF CRUSTAL DEFORMATION

Folding and faulting of the Earth's crust produces mountains, such as the Rocky Mountains, Appalachian Mountains, Himalayas, and Alps. Jointing, by contrast, plays only a minor role in shaping the terrestrial surface.

Folding. Compression of the Earth's surface results in the formation of mountain ranges in which rocks are extensively folded. Erosion of a sequence of sedimentary rocks containing beds, some of which are easily eroded and others of which are resistant to erosion, produces numerous parallel ridges and valleys that are oriented parallel to the mountain belt in which they are found. Such ridges and valleys are well developed in the Valley and Ridge Province in the Appalachians and in the eastern part of the northern Canadian Rockies. Ridges may be underlain by anticlines, synclines, or homoclines (Figure 34). Valleys also may be underlain by anticlines, synclines, and homoclinal ridges or homoclines (Figure 34). Anticlinal, synclinal, and homoclinal ridges have resistant sedimentary layers beneath them, whereas the corresponding types of valleys are underlain by less resistant layers.

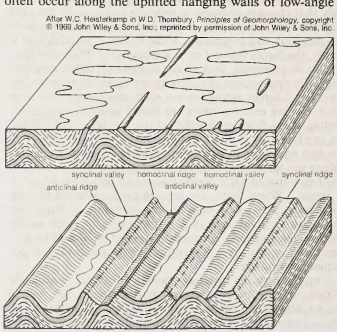
Faulting. This type of fracturing results either from the compression or the extension of the Earth's crust, and it may lead to the formation either of parallel mountain ranges and valleys or of one or two very elongated valleys bordered by broad uplifts instead of mountain ranges.

Extension of the Earth's crust may result in the formation of numerous parallel ridges and valleys oriented perpendicular to the direction of extension. Such extension may occur parallel to the coastline of a continent and cause parallel sets of normal faults to form. Generally ridges develop along the uplifted block of the footwall of a normal fault, while valleys form along the dropped block of its hanging wall. The Basin and Range Province of the western United States is a region in which ridges have been uplifted and valleys have been dropped along normal faults. This region may have been subjected to extension as a result of the formation of a broad arch underlying the Basin and Range Province. Stretching of the Earth's crust at the top of the arch may have created the extension, which in turn produced the normal faulting.

Formation of long valleys underlain by grabens

Normal faults bordering grabens are formed by crustal extension, and this results in the formation of long valleys underlain by the hanging walls of the normal faults that have been dropped relative to the footwalls of the uplifted blocks bordering the valleys. The East African Rift System is an excellent example of a series of valleys underlain by grabens formed by crustal extension.

Reverse faults are common in mountain belts and are partly responsible for the formation of the belts. Ridges often occur along the uplifted hanging walls of low-angle



After W.C. Hesterkamp in W.D. Thornbury, *Principles of Geomorphology*, copyright © 1969 John Wiley & Sons, Inc., reprinted by permission of John Wiley & Sons, Inc.

Figure 34: The topographic expressions of eroded anticlines and synclines.

reverse faults in mountains such as the northern Rockies in western Canada and the northwestern United States. Valleys develop between these uplifted blocks in part because the rocks along low-angle reverse faults are brecciated during faulting and are thus less resistant to erosion.

High-angle reverse faults border many of the mountain ranges of the central and southern Rockies. There, the uplifted hanging walls of the faults form large mountains that are somewhat isolated from one another and often are not parallel to each other. This contrasts with the mountains of the northern Rockies, which tend to parallel each other and are connected together. The mountains of the central and southern Rockies are separated by broad basins in which thick sequences of sedimentary rocks have been deposited as a result of erosion of the adjacent mountain ranges.

Many strike-slip faults contain a zone several kilometres wide along which rocks are intensely brecciated and sheared. Erosion of these brecciated and sheared zones commonly results in the formation of an elongated valley paralleling the underlying fault zone. Such valleys are common along the San Andreas Fault in California, particularly in the region south of San Francisco.

Formation of joints. Because joints are areas of weakness in sedimentary rocks, erosion may occur more readily along regions in which joints are closely spaced. Consequently, streams in some areas of jointed rocks tend to parallel one or more joint sets.

THE SURFACE OF THE EARTH AS A MOSAIC OF PLATES

When earthquake epicentres are plotted on a map of the world, it is apparent that most of them are confined to relatively narrow bands rather than being uniformly distributed over the Earth's surface. The narrow bands along which earthquakes occur are located at the margins of large plates that are moving relative to each other. As previously noted, these plates consist of the entire crust of the Earth plus a significant portion of the upper mantle. About a dozen plates make up the Earth's surface. (For a detailed discussion of this subject and a map of the principal plates, see the article PLATE TECTONICS.)

Size of the plates

Geometry and rate of plate movement. The plates range in size from about 400 by 2,500 kilometres to 10,000 by 10,000 kilometres. Some Earth scientists believe that the plates are equal in thickness to the lithosphere (see above). If this is so, the plates would be between 70 and 225 kilometres thick. Other investigators hold that the lithosphere moves along with the asthenosphere (the relatively plastic layer underlying the lithosphere) and perhaps the top of the mesosphere as well. In this case, the plates could be as much as 650 kilometres in thickness. Because earthquakes extend to depths of 650 to 700 kilometres, it is clear that plate movements extend to at least that depth.

There are two types of plates, continental and oceanic. An example of a continental plate is the North American Plate, which includes North America as well as the oceanic crust between it and a portion of the Mid-Atlantic Ridge. Oceanic plates are made up primarily of oceanic crust. A plate of this type is exemplified by the Pacific Plate, which extends from the East Pacific Rise to the trenches bordering the western part of the Pacific Basin.

The velocities of plates vary with distance from the spreading pole of the plates. The spreading pole is an imaginary axis about which one plate moves relative to another. At any one time, the velocity of plates will be at a maximum 90° (great circle degrees) away from the spreading pole and will be at a minimum at or near the pole of spreading. At present there are significant differences between the maximum spreading velocities of oceanic plates and those of continental plates. In general, the maximum velocity of spreading of the oceanic plates is two to four times that of the latter. The maximum spreading velocity of the Pacific Plate relative to the East Pacific Rise, for example, is about eight centimetres per year, while that of the North American Plate relative to the Mid-Atlantic Ridge is roughly two centimetres per year.

Types of plate boundaries. As explained earlier, there are three types of plate boundaries: divergent, convergent, and transform (or strike-slip). Plates move in opposite directions along the same line at divergent plate boundaries. These boundaries are located at mid-oceanic ridges (spreading ridges) such as the Mid-Atlantic Ridge and the East Pacific Rise. Ridges of this kind are part of a worldwide system of ridges that generally occur at the centres of ocean basins. The system forms an undersea mountain chain measuring more than 60,000 kilometres in length. Most or perhaps all of the component ridges originally formed beneath continents at the time when continental fragmentation began. Take for example the Mid-Atlantic Ridge, which emerged about 200,000,000 years ago when North America and Eurasia began to separate from Gondwanaland (which consisted roughly of present-day South America, Africa, India, Australia, and Antarctica).

Divergent plate boundaries

The second type of plate boundary, the convergent boundary, forms where two plates move toward each other along the same line. Such boundaries are located either at or near the margin of a continent where the seafloor moves under the continental margin along a subduction zone, or within a continent where two continents collide. Oceanic trenches typically occur along subduction zones. The trenches bordering the Pacific Basin (*e.g.*, the Peru-Chile Trench and the Aleutian Trench) are located along such convergent plate boundaries. Intracontinental mountain ranges occur along convergent plate boundaries where two continents are colliding. The Himalayas are situated along a plate boundary of this sort.

Convergent plate boundaries

The third type of plate boundary, the transform variety, forms where two plates are moving in opposite directions but not along the same line. Transform boundaries are located along transform faults, which are a type of strike-slip fault, as, for example, the San Andreas Fault. Such faults develop where the motion of plates is transformed from one type of motion to another. The most common type of transform fault offsets spreading ridges (Figure 35). Here, the motion is transformed from along the fault. The actual motion along such a transform fault is exactly the opposite of the apparent offset of the spreading ridge. Transform faults also may connect two trenches, a ridge and a trench, a mountain and a ridge, and so forth.

Transform plate boundaries

Activity along plate boundaries. In addition to earthquakes, most of the faulting, folding, igneous activity, metamorphism, and mountain building on the Earth occurs at or near plate boundaries.

Earthquakes. Most Earth scientists believe that virtually all naturally produced earthquakes occur as a result of movements along faults. The reason that the majority of such earthquakes occur along plate boundaries is that most faults lie along those boundaries.

Faulting. Normal faults are very common along divergent plate boundaries because these are regions of crustal stretching (tension). The oceanic crust along the crests of mid-oceanic ridges is cut by numerous normal faults trending parallel to the ridge crests (Figure 36). Most of

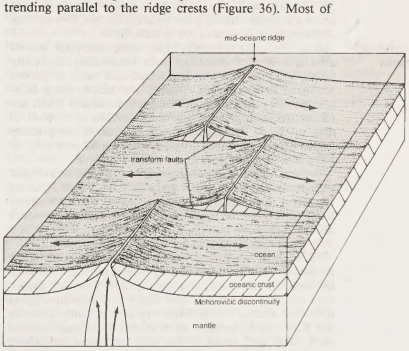


Figure 35: Two transform faults offsetting a mid-oceanic ridge.

these faults, however, are located below sea level and can be found only by submarine depth profiling.

Normal faults are formed in continental crust along divergent plate boundaries when two continents begin to separate from each other. Continental crust is uplifted, stretched, and faulted over the convection current that caused the continents to separate. Such faults formed along the margins of the Atlantic and Indian oceans when the continents began to pull apart from each other approximately 200,000,000 years ago. The majority of these faults are covered either by younger sedimentary deposits that now underlie the continental shelf or by deposits of the coastal plains bordering the oceans.

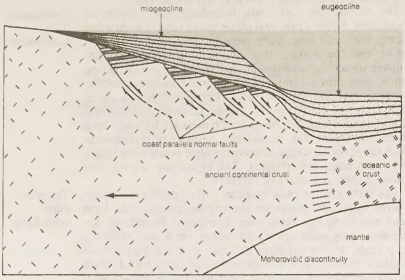


Figure 36: Cross section of a continental margin adjacent to a divergent plate boundary.

Reverse faults are very common along convergent plate boundaries where rocks are undergoing compression. They are found in areas where the seafloor is moving beneath continental margins along subduction zones and in areas where two continents collide. Reverse faults may be produced beneath a plate margin in a subduction zone (Figure 37, top). They also may develop above the plate margin along a wide belt on the continental margin parallel to the flanks of a mountain range (Figure 37, bottom). Reverse faults are common along the Pacific margin of continents. They also may be produced both in front of and beneath an overriding continent where two continents push together (Figure 38). Many are found in the Himalayas and the southern Appalachians. The Himalayas are at the present time the site of a collision between India and Asia, while the southern Appalachians were the site of a continent–continent collision some 250,000,000 to 350,000,000 years ago. Normal and strike-slip faults also occur along convergent plate boundaries, but they are less common than reverse faults in such areas.

Strike-slip faults

Strike-slip faults frequently occur along transform boundaries and, in fact, constitute such boundaries. Strike-slip faults, however, may also be formed as a result of stresses produced in the Earth's crust by movement along adjacent primary strike-slip faults. Such strike-slip faults are termed secondary faults and may develop as a result of drag produced by frictional resistance along the primary faults. Secondary normal and reverse faults also may be formed by drag along primary strike-slip faults.

Folding. Folds are produced principally at convergent plate boundaries, but minor secondary folds develop along transform plate boundaries due to drag. They are found along plate boundaries adjacent to subduction zones where oceanic plates are moving beneath continental plates and where one continent collides with another continent.

Folds along subduction zones are attributable to shear stresses created by drag resulting from frictional resistance between the two plates (Figure 37, top). They are associated spatially with the reverse faults in subduction zones discussed above. The sedimentary and volcanic rocks in the Klamath Mountains of northern California contain such folds and faults. Folds also are produced above subduction zones as a result of the compression of plate

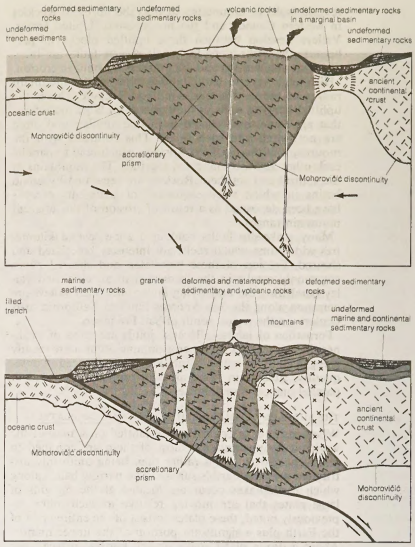


Figure 37: Cross section of a convergent plate boundary involving a collision between a continental plate and an oceanic plate in the vicinity of (top) an island arc and (bottom) a mountain arc.

margins by the descending oceanic slab. These types of folds and reverse faults occur in the northern Rockies.

Folds that occur where one continent collides with another may be the product of the shearing or compression caused by the shoving of one continent over another continent. Folds just above or just below such a plate margin along a subhorizontal plate boundary are formed by the drag associated with the shearing motion of the continents as they slide past each other. These folds are common in the Himalayas and in the central part of the southern Appalachians. Folds also may develop in front of a continent that is colliding with another continent. Such folds

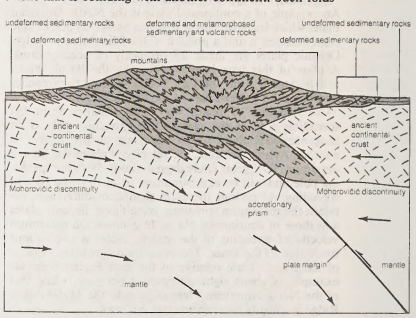


Figure 38: Cross section of a convergent plate boundary involving a collision between two continental plates near a Himalaya-type mountain chain.

are produced by the compressive force exerted in front of one continent that is overriding another continent. An example of folds of this variety is found in the Valley and Ridge Province in the southern Appalachians.

Formation of igneous rocks. Igneous rocks are formed by the crystallization of magma. Extrusive igneous rocks (volcanic rocks) are produced by the crystallization of magmas at the surface. Intrusive igneous rocks include those crystallized at shallow depth (hypabyssal igneous rocks), typically as dikes and sills, and those crystallized at medium to great depths (plutonic igneous rocks).

Igneous rocks are commonly found along both divergent and convergent plate boundaries. Those along divergent plate boundaries include volcanic, hypabyssal, and plutonic rocks. They form at spreading ridges and are part of the oceanic crust produced at these sites as plates move apart. Ophiolite sequences found on the continents are thought to be such oceanic crust brought to the Earth's surface by tectonic forces. The uppermost layer of igneous rocks in ophiolite sequences formed at divergent plate boundaries is composed of a basalt, a fine-grained, extrusive igneous rock of mafic composition (*i.e.,* rich in iron and magnesium). The next layer of igneous rock in ophiolite sequences is diabase, a mafic hypabyssal intrusive igneous rock. This type of diabase generally occurs as narrow dikes, more or less vertical intrusions that cut across sedimentary layering. These dikes are thought to have been the conduits that fed the basaltic flows in the overlying layer of pillow basalts. A layer of gabbro underlies the diabase in an ophiolite sequence. Gabbro is a mafic plutonic intrusive igneous rock that probably formed in a magma chamber beneath both the dikes and flows. A layer rich in the minerals olivine or serpentine underlies the gabbro and is thought to represent the mantle of the Earth. The Mohorovičič discontinuity separates the basalt, diabase, and gabbro of the oceanic crust from the underlying mantle rich in olivine and pyroxene.

Igneous activity at the spreading ridges

There is considerable variation in the amount of igneous activity along divergent plate boundaries at spreading ridges. Igneous activity is appreciably more intense in the vicinity of hot spots than elsewhere along divergent plate boundaries. A hot spot is an area in which igneous activity, particularly volcanism, is greater than average. Some hot spots are thought to be underlain by a mantle plume (a column of hot, partially molten rock emanating from the mantle). Iceland is an example of such a hot spot.

Hot spots

The igneous rocks formed along divergent plate boundaries also include those emplaced on or within continental crust. During the initial stage of the separation of two continents, continental crust is uplifted and faulted beneath the rising convection currents that cause the continents to pull apart. Basaltic magmas rise along such fractures and crystallize as volcanic rocks from flows or as dikes and sills if they form shallow intrusions. Such igneous rocks are often preserved in fault basins paralleling the continental margin and occur along with thick sequences of sedimentary rocks (Figure 36). An example of igneous and sedimentary rocks preserved in such a fault basin are the deposits of the Newark Series and rocks of a similar age found in the Newark and Gettysburg basins in the northeastern United States. These deposits and the fault basins in which they occur were formed when North America and Asia began to separate from Gondwanaland.

The origin of igneous rocks at divergent plate boundaries is presumably related to convection currents thought to rise under spreading ridges at the boundaries. As mantle material (composed of the rock peridotite) rises under the ridges, pressure is released on this material. Because the material is already hot and because a decrease in pressure tends to lower the melting temperature of rock, the rising peridotite begins to melt, or if it is already partly molten, the percentage of melt increases. When the amount of peridotite melt reaches 25 percent or so, the melt starts to separate from the solid, and the magma rises to form the igneous rocks at the divergent plate boundaries.

Igneous rocks also occur at convergent plate boundaries, both adjacent to subduction zones where oceanic plates move beneath continental plates and where two continental plates collide. Volcanic and intrusive varieties form along the so-called ring of fire bordering the Pacific Basin. Mt. St. Helens, Mt. Fuji, and the volcanoes of the Andes Mountains are examples of volcanoes that occur adjacent to subducting plate margins. The volcanic rocks in such areas include basalt, rhyolite (a sialic volcanic rock, rich in silicon and aluminum), and andesite (a rock intermediate in composition between basalt and rhyolite). Andesitic volcanic rocks are thought to form due to the partial melting of mantle material in the continental plate above the plate margin in the presence of a significant amount of water (Figure 37, top). The water may have been released from the oceanic plate that is being subducted beneath the continental plate. Magmas that form the volcanic rocks adjacent to subduction zones also may be produced by the frictional (shear-strain) heating of the rocks along the plate margin. This heating results from the frictional resistance generated along the plate margin as the oceanic plate moves past the continental plate.

Volcanic and intrusive igneous rocks in regions of convergence

Plutonic igneous rocks also form adjacent to subduction zones where an oceanic plate moves under a continental plate. Some of these plutonic rocks (gabbros and diorites) comprise the same magmas that formed the volcanic rocks of basaltic-to-andesitic composition. Intrusions of granite (sialic intrusive igneous rocks), however, often form very large bodies called batholiths. Granitic batholiths are thought by most, but not all, geologists to have been produced by the partial melting of material in the lower crust along subduction zones at convergent plate boundaries (Figure 37, bottom). Studies of the isotopic composition of strontium from batholiths found adjacent to subduction zones indicate that some batholiths (*e.g.,* the Sierra Nevada Batholith) were formed by the partial melting of sedimentary and volcanic rocks that were added onto the continental margin by accretion within the past 1,000,000,000 years, while others (*e.g.,* the Idaho Batholith) were produced by partial melting of rather ancient continental material more than 1,500,000,000 years old. After such a partial melt forms, it rises in the crust and provides materials for batholiths and volcanic flows or volcanic ash.

Origin of granitic batholiths

Igneous rocks found adjacent to plate margins where two continents converge tend to be of sialic composition and comprise granitic intrusions (including granite batholiths) and rhyolitic extrusions (both volcanic flows and volcanic ash). These sialic magmas are thought to have been formed primarily by the partial melting of crustal rocks near subducting plate margins. The heat necessary to partially melt crustal rocks along plate margins may have come from frictional heating as well as from hot rocks in the lower crust of the continental plate overlying the plate margin. This situation prevails in the Himalayas.

Metamorphism. Metamorphism requires heat and/or pressure to transform a sedimentary or an igneous rock into a metamorphic rock. Such a change involves an increase in grain size, formation of new minerals, gain or loss of water, and/or an increase in the degree of bonding of the mineral grains. Metamorphism can be either of two types: regional or contact. Regional metamorphism involves areas of very large extent (a few thousand to hundreds of thousands of square kilometres), while contact metamorphism is restricted to areas adjacent to igneous intrusions such as batholiths. Contact metamorphism can occur anywhere igneous intrusions occur. Regional metamorphism occurs principally along divergent and convergent plate boundaries. Metamorphism along divergent plate boundaries results in the conversion of basalts, diabases, or gabbros to greenstones or amphibolites. The greenstones are produced as a result of the heat furnished by the upwelling convection currents beneath the boundaries.

Regional and contact metamorphism

Regional metamorphism also may take place at convergent plate boundaries either where an oceanic plate moves under a continental plate along a subduction zone or where two continents collide. Regional metamorphism adjacent to subduction zones may occur beneath the plate margin in the underlying oceanic plate or above the plate margin in the overlying continental plate.

In cases where regional metamorphism occurs on the descending edge of an oceanic plate, sedimentary and volcanic rocks such as shales, sandstones, limestones, and basalts are converted into schists, quartzites, marbles,

greenstones, amphibolites, blueschists, and other metamorphic rocks (Figure 37, top). Some of these rocks (*e.g.*, the blueschists) show evidence of having been metamorphosed under relatively high pressures and low-to-moderate temperatures. Such conditions prevail wherever sedimentary and volcanic rocks on oceanic crust are carried rapidly beneath a continental plate. An increase in temperature in these areas is produced by frictional heating caused by the drag of one plate past the other and by heat from the overlying continental plate. An increase in pressure within the rocks is caused by the weight of the overlying rocks in the continental plate and stresses (overpressures) developed during the shearing of rocks along the plate margin. Distinct manifestations of metamorphism beneath plate margins along subduction zones are evident in many rocks bordering the Pacific Basin—*e.g.*, in the blueschists of the Franciscan Formation of the California Coast Ranges.

Regional metamorphism in the continental plate above the plate margin near subduction zones typically involves conditions of relatively high temperatures and low-to-medium pressures. Metamorphic rocks produced under such conditions include schists, marbles, quartzites, and amphibolites. The causes of the pressure increase during the metamorphism in areas of collision between two continents are both the weight of the continent above the plate margin and the stresses built up during the shearing of rocks along plate margins (Figure 38). Examples of areas affected by this type of metamorphism include the Himalayas and the Grenville Province in Canada.

Mountain building. Mountains are areas that have been uplifted as a result of either the thickening of the Earth's crust or the heating of the crust and upper mantle that causes an expansion of these layers. Mountain building occurs along both divergent and convergent plate boundaries.

The mid-oceanic ridge

The mid-oceanic ridge system is located along divergent plate boundaries. This system is an undersea chain of mountains that extends up to one-third the width of the oceans in which it is found. The crust under the ridges has about the same thickness as the crust in the adjacent ocean basins. Thus, the uplift of the mid-oceanic ridges is not due to a thickening of the crust but rather to the heating of the crust and upper mantle beneath the ridges. This heating is presumably caused by the rising convection currents underlying mid-oceanic ridges.

Mountains also are formed in areas underlain by continental crust along divergent plate boundaries where two continents are just beginning to pull apart from each other. Such mountains form because of uplift caused by heating of the crust and mantle under the continents. This heating is attributable to the formation of a convection current beneath the continents, the current being the force that causes the landmasses to start separating. The mountains presently bordering the Red Sea in Africa and Saudi Arabia are representative of those formed in this fashion.

The mountains bordering the Pacific Basin formed along convergent plate boundaries where an oceanic plate moved under a continental plate along a zone of subduction. Compression of the continental plate by the descending oceanic plate resulted in the thickening of the crust, especially in the Rockies and Andes, and this crustal thickening produced the mountains. In some places the crust has been thickened from an average of 35 kilometres to as much as 70 kilometres and in the process produced the uplift necessary for mountain building.

Mountains that formed as a result of continent–continent collisions include the Appalachians and the Himalayas, along with the adjacent Plateau of Tibet. These mountains may have been created by the movement of one continental margin beneath another, which caused the thickness of the crust in the affected areas to double. For example, the thickness of the crust in southern India measures roughly 35 kilometres, but its thickness under the Plateau of Tibet north of the Himalayas is about 70 kilometres.

Intraplate activity. Although most tectonic activity takes place along plate boundaries, some occurs within plates. Much of this intraplate activity occurs along aulacogens and plume traces. An aulacogen can be defined as a failed third arm of a spreading ridge. Aulacogens commonly intersect a divergent plate margin at an angle of about 120°, and they form just as two continents begin to separate. They intersect plate margins at the site of the upwelling of a mantle plume (Figure 40). They are regions of tension within a plate and are commonly manifested in the form of grabens. Aulacogens often are filled with thick sequences of sedimentary and volcanic rocks, and volcanoes are usually located along them. The volcanic activity develops in response to upwelling mantle material beneath aulacogens. This material, however, does not rise with enough velocity to cause plate separation, only plate stretching. As apparent from Figure 39, the East African Rift System is an excellent example of an active aulacogen. This rift system intersects the margin of the African Plate at a plume located at the southern end of the Red Sea in the vicinity of the Afar Triangle.

When a plate moves over a mantle plume, rocks that may be part of a volcano are carried away from the plume in the direction of plate motion. Plumes located within a plate produce a chain of volcanic islands and seamounts (submarine volcanoes) extending in one direction away from the plume. The age of the volcanic rocks and of the volcanoes generally increases away from the plume. Such a chain of volcanic islands and seamounts is called an aseismic ridge (a ridge without earthquakes), or plume trace.

Aulacogens

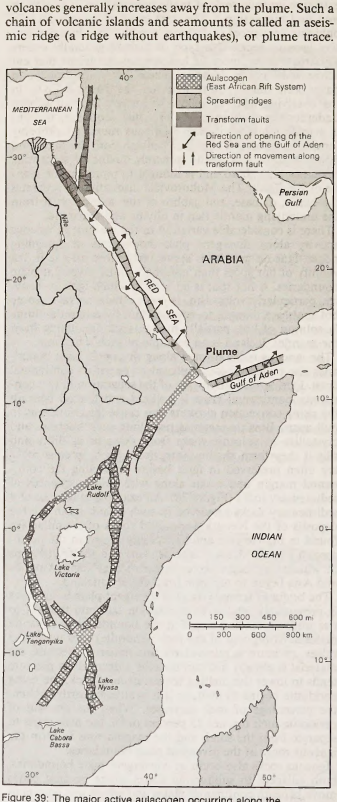


Figure 39: The major active aulacogen occurring along the East African Rift System.

Plume traces

The Hawaiian and Emperor Seamount chain is representative of this kind of structure. If a plume is located on a spreading ridge, as many are, two plume traces will form and extend in opposite directions away from the spreading ridge (Figure 40). Examples of such plume traces include a chain of volcanic islands extending westward away from the East Pacific Rise through Tuamoto Island and another volcanic island chain extending eastward through Easter Island away from the rise. Much of the volcanic activity along a plume trace formed by a plume on a spreading ridge is concentrated at the crest of the ridge, but a significant amount of volcanic activity along many such plume traces occurs within a plate after the plate has moved off the plume. Volcanic activity on Ascension Island, Easter Island, Pitcairn Island, and Tristan da Cunha is illustrative of this type of intraplate volcanism along a plume trace ending at a plate margin.

Convection theory

The cause of plate motions. Plates are thought to move in response to the movement of convection currents. These currents cause a vertical transfer of heat through a mass movement of material. In the Earth, the material that is moving vertically is mantle material, which contains about 5 percent magma between crystalline solids. This amount of liquid allows the mantle material to move more easily than if it were entirely solid.

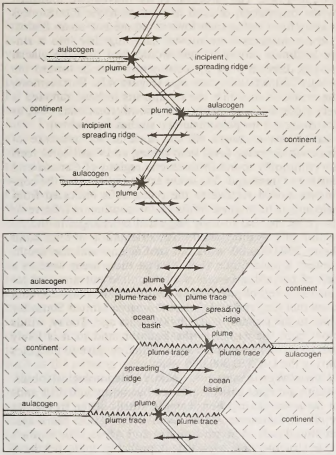


Figure 40: The development of aulacogens, plumes, and plume traces during the separation of two continents. Here, the process is shown in two stages.

The configuration of convection currents within the mantle is controversial. Some Earth scientists believe that convection within the mantle resembles convection cells observed within Newtonian liquids. Because the mantle is largely solid, however, it is unlikely that this type of convection would occur. Convection within a plastic solid would instead be expected to result in rising hot dikes. Such dikes are presumed to move upward under the influence of gravity and as they rise push plates apart, thereby providing the principal mechanism for plate movement. Some plates may move partially as a result of the pull of the relatively cool descending slab as it sinks down under the influence of gravity. Slab pull, however, cannot be the dominant force driving the plates because there are several very large plates that do not have a descending slab attached to them. Such plates include the North American and Eurasian plates.

To determine the configuration of the convection currents that move the plates, it is necessary to consider changes in the size of the plates. At the present time the plates bordering the Atlantic and Indian oceans (the North American, South American, Eurasian, Indian-Australian, African, and Antarctic plates) are increasing in area and volume as the continents drift away from the Mid-Atlantic and Mid-Indian ridges. As the continents move, the area of the Atlantic and Indian ocean basins increases. At the same time the plates bordering the Pacific Ocean (*e.g.*, the Pacific, Nazca, and Juan de Fuca plates) are decreasing in area and volume as the continents move toward the centre of the Pacific Basin. Hence, the area of the Pacific Basin is decreasing. In order for the plates to change in size, there must be a transfer of material from some plates to others. It is likely that material is transferred from the Pacific Plate to the Indian Plate by a movement of material from the subduction zone north of New Zealand to the Mid-Indian Ridge and from the subduction zone west of South America on the eastern part of the Nazca Plate to the Mid-Atlantic Ridge (Figure 41). Many convection models do not show such a transfer of material, but this transfer is necessary if plates do change in size as indicated above.

Convection currents are thought to rise unusually high under mid-oceanic ridges, and these are regions that exhibit a great deal of volcanic activity. Both these features would be expected to occur over a rising convection current. As the plates are forced apart by such currents, new oceanic crust is created at the ridges, and the ridges move away from the continents. For example, the Mid-Atlantic Ridge is moving westward away from Africa, the Southwest Indian Ridge southeastward from the continent, and the Mid-Indian Ridge northeastward from it. Geometry requires that as these ridges move away from Africa, they increase in length. Plumes located on these ridges move radially away from Africa. Because the ridges are increasing in length, the distance between the plumes located on them increases as well.

Energy sources for convection. Measurements of temperature in deep mines and oil wells indicate that there is a substantial temperature increase with depth. In part, this is due to the weight of overlying rocks. Such temperature increases are not accompanied by a change in heat content and are referred to as adiabatic temperature increases. Convection within the Earth occurs only if the temperature increase with depth exceeds the adiabatic temperature increase. Temperature increases of this kind are called super-adiabatic temperature increases.

Super-adiabatic temperature increases

The Earth has within it a super-adiabatic temperature increase. The energy sources causing this temperature increase include (1) the decay of long-lived radioactive (unstable) isotopes, such as uranium-235, uranium-238, and potassium-40; (2) the change of part of the Earth's core from a liquid to a solid; (3) the original heat from the time when the Earth formed, including heat produced by the impact of meteorites during the growth of the planet; and (4) the flexing of the Earth during tidal interaction between the Earth, Sun, and Moon (especially if the Earth and Moon were significantly closer together in the past).

Evidence for polar wandering, continental drift, and seafloor spreading. Polar wandering, continental drift, and seafloor spreading are all consequences of plate movements. Polar wandering is the movement of a continent relative to the rotational poles or spin axis of the Earth. Continental drift is the movement of one continent relative to another continent, and seafloor spreading is the movement of one block of seafloor relative to another block of seafloor. Evidence for both polar wandering and continental drift comes from matching continental outlines, paleoclimatology, paleontology, stratigraphy, structural geology, and paleomagnetism. The concept of seafloor spreading is supported by the age of volcanic islands and the age of the oldest sediments on the seafloor, as well as by the study of the magnetism of the seafloor.

Matching continental outlines. Early geographers making maps of the South Atlantic Ocean were probably the first to notice the similarity of the outlines of South America and Africa and to wonder if these two continents might have been together at one time. It was not until

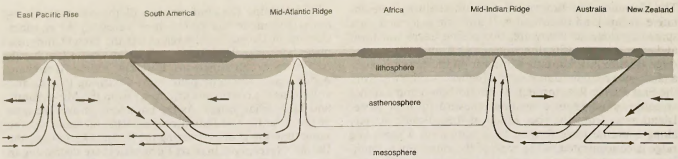


Figure 41: Schematic cross section showing one of various possible models for convection within the Earth.

the early 20th century, however, that Alfred Wegener of Germany used the geography of the Atlantic coastlines, along with geologic and paleontological data, to suggest that all the continents were once connected. In his major work *Die Entstehung der Kontinente und Ozeane* (1915; *The Origin of Continents and Oceans*), he postulated that a single supercontinent called Pangaea existed some 320,000,000 to 286,000,000 years ago and that its subsequent breakup gave rise to the present-day continents.

Wegener's hypothesis

Actually, the fit of the coastlines of South America and Africa is only fair. These continents fit together very well, however, if they are matched at the 1,000-metre depth contour. In recent years computers have been used to match the outlines of all continents bordering the Atlantic Ocean (Figure 42). Such computer fits are made so that continental outlines at different depths are matched with a minimum area of overlap and a minimum area of gap between the continents. Similar fits of continents may be made around the Indian Ocean as well.

Paleoclimatological data. Alfred Wegener was a meteorologist who was particularly interested in the study of ancient climates (paleoclimatology). He noticed that sedimentary rocks indicative of having been deposited in warm climates are now found in cold climates and vice versa. He felt that these observations were best explained if the continents had moved relative to each other and if the poles had moved relative to those continents.

From Scientific American (April 1968), p. 52



Figure 42: Computer-generated "best fit" of the continents bordering the Atlantic Ocean, as proposed by the British geophysicists E.C. Bullard, J.E. Everett, and A.G. Smith. The fit was made at the 1,000-metre (500-fathom) submarine depth contour. The matching was done in such a way that the area of the overlaps (in black) of the continental margins equals the area of the gaps (in white) between them.

Glaciation is indicated by the presence of till containing grooved, striated, and faceted pebbles, along with cobbles and boulders that lie on grooved, striated, and polished bedrock. Till deposited by continental ice sheets 420,000,000 years ago is found in the Sahara, and similar unconsolidated sediments dating back 280,000,000 years occur near the present-day Equator in southern India and in central Africa. The direction of ice movement, indicated by the orientation of such features as roche moutonnées is away from rather than toward the present ocean basins. If the continents were moved back together into a single landmass, however, one would find that the glacial ice was directed from the continental interior toward the oceans 280,000,000 years ago (Figure 43). Furthermore, the centre from which the ice appears to have radiated is very close to the location of the south rotational pole, as determined from paleomagnetic studies (see below).

Indications of warm climate such as the presence of thick salt deposits, fossil sand dunes, and fossil coral reefs are commonly found in the arctic regions of Canada and northern Greenland. These are easily explained if the continents moved relative to the poles, as is indicated from paleomagnetic studies. Shifts of the continents relative to the poles are also strongly suggested by changes in the direction of prevailing winds. The orientation of fossil sand dunes 250,000,000 years old in England and the southwestern United States indicate that these areas were in a belt of prevailing easterlies at that time but are now located in a belt of prevailing westerlies. Paleomagnetic data show that England and the southwestern United States lay close to the Equator 250,000,000 years ago. Today, winds blow from the east near the Equator, and so it is likely that those two geographic areas shifted relative to the poles during the last 250,000,000 years.

Paleontological research. Many paleontological observations are most easily explained by continental drift and/or polar wandering. For example, fossils of the freshwater reptile *Mesosaurus,* which lived 270,000,000 years ago, are found only in South America and Africa. The reptile's bone structure was such that it probably could not have swum across the Atlantic, but if South America and Africa had been connected then, *Mesosaurus* could easily have moved from one continent to another.

Distribution of fossils

Terrestrial reptiles that lived 225,000,000 years ago in Antarctica closely resemble those of the same age found in Africa. These two areas are not connected at the present time, but they may have been joined together 225,000,000 years ago. Furthermore, fossils from plants belonging to the genus *Glossopteris* and related genera that lived in South America, Africa, Australia, India, and Antarctica 270,000,000 years ago are remarkably similar, leading to the hypothesis that these landmasses once constituted the supercontinent Gondwanaland (see above).

Stratigraphic findings. The sequence of layered rocks on the landmasses that constituted Gondwanaland is strikingly similar for those time periods when the landmasses are believed to have been together. In these areas glacial deposits, for example, are overlain by coal-bearing shales and sandstones containing fossils of *Glossopteris* and *Mesosaurus,* which are in turn overlain by thick sequences of mafic (basaltic) volcanic rocks.

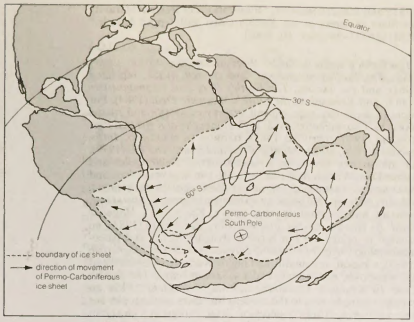


Figure 43: Paleogeographic map of the continents during the Late Carboniferous and Early Permian periods showing the inferred distribution of continental ice sheets.

Results of structural geologic studies. Deformed rocks typically occur in long, linear belts along the trend of present or past mountain belts. In many parts of the world, mountain belts end rather abruptly at the edges of continents and commonly occur in pairs located on opposite sides of an ocean. The Appalachian mountain belt, for example, extends from Georgia and Alabama through southeastern Pennsylvania, New England, and the Maritime Provinces of Canada to Newfoundland. The Caledonian mountain belt, which, like the Appalachians, formed about 450,000,000 to 350,000,000 years ago, extends from the island of Spitsbergen north of Norway through Norway, Scotland, and England to Ireland. Another part of the Caledonian belt is located on the east coast of Greenland. When all the continents are put back together in a single landmass, the Appalachian and Caledonian mountains are connected into a single mountain range (Figure 43).

Similarities between mountain belts

Paleomagnetism. All rocks contain what is referred to as natural remanent magnetism. This property is due to magnetism acquired by iron-bearing minerals during the cooling and crystallization of an igneous rock, to the formation of iron-bearing minerals at low temperatures during the weathering of previously formed minerals, to the rotation of iron-bearing minerals during their deposition in a sedimentary rock, or to post-depositional processes such as the striking of rocks by lightning.

Natural remanent magnetism

The direction and inclination of the magnetic field of rocks of different ages have been measured from rock samples collected from all over the world, and this information can be used to ascertain the location of the Earth's magnetic pole at the time when those rocks were formed. The direction of the magnetic field of a given rock sample indicates the direction in which the magnetic pole of the Earth lay when the rock formed, while the inclination of the magnetic field of the rock indicates how far away from the collection site the magnetic pole was located. For example, if the inclination of the magnetic field is nearly horizontal, the magnetic pole of the Earth must have been 90 great circle degrees away from the collection site because the site was near the magnetic equator. On the other hand, if the inclination of the magnetic field of a rock is vertical, the collection site would have been located at or near the Earth's magnetic pole at the time of rock formation. It is assumed that if enough rock samples of a given age are averaged together, the average position of the magnetic pole will be the same as the average position of the Earth's rotational pole. Thus, paleomagnetic poles provide the location of the planet's rotational pole.

If paleomagnetic poles for a given continent are plotted on a map of the world, it is found that they tend to lie along lines called apparent polar wandering (APW) paths (Figure 44). Evidently Europe drifted eastward relative to North America sometime during the past 550,000,000 years. If Europe were moved westward back to the position that it occupied before the continents began to separate from one another (as indicated by matching continental outlines across the Atlantic), the APW path for Europe from 200,000,000 to 400,000,000 years ago agrees with the equivalent APW path for North America. Thus, Europe and North America were together in the arrangement indicated by matching of continental outlines during that time interval.

Age of oceanic islands. One of the earliest indications that some blocks of seafloor are moving relative to other such blocks was that while the age of young volcanic islands or seamounts shows no relationship to distance from spreading mid-oceanic ridges, old features of this sort are never found near the ridges. This relationship can be readily explained if new seafloor is created at the crests of mid-oceanic ridges and with time this seafloor, along with the volcanic islands, is carried off the ridges as the seafloor spreads. In this case, there can be no old volcanic islands or seamounts near the ridges because the oceanic crust on which volcanic islands sit is relatively young near the ridges.

Evidence for the creation of new oceanic crust at the mid-oceanic ridges

Age of the oldest marine sediments. Samples of sediments have been recovered as a result of deep drilling and piston coring from many parts of the world's ocean basins. These sediments have been dated by means of the fossils they contain. Such studies have shown that no old sediments are present (at least in the areas sampled) near mid-oceanic ridges. Moreover, the studies also have revealed that the age of the oldest sediments in the ocean basins generally increases away from the ridges. This relationship is exactly what one would expect if new oceanic crust is being formed at the ridges and is being carried away from them as the seafloor spreads.

Marine magnetics. The most convincing evidence supporting the concept of seafloor spreading comes from the study of the magnetism of the ocean floor. For a number of years, measurements have been made in the oceans of the total intensity of the magnetic field of the Earth using instruments called magnetometers, which are generally towed behind ships (see above). These measurements have shown that there are numerous magnetic anomalies (deviations from the average intensity of the Earth's magnetic field) in the ocean basins. Such anomalies are positive if

From *Earth*, 3/E by Frank Press and Raymond Siever, copyright © 1974, 1978, 1984 W.H. Freeman and Company; used by permission

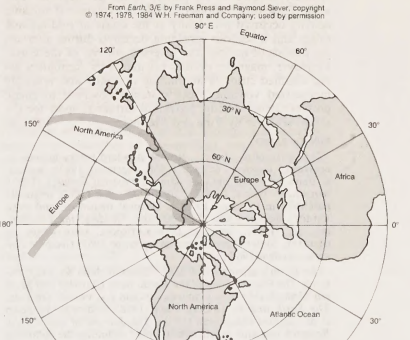


Figure 44: Comparison of apparent polar wandering (APW) paths of North America and Europe. The fact that the APW paths do not coincide indicates that these continents have moved with respect to each other.

the magnetic field is greater than average, and they are negative if the field is less than average.

Magnetic anomalies along the mid-oceanic ridges

Magnetic surveys over the crests of mid-oceanic ridges indicate that magnetic anomalies occur in long bands paralleling the axis of the ridges and that the anomalies are symmetrical with respect to the ridges.

Rocks that are relatively young and that have a magnetism roughly parallel to the present-day magnetic field of the Earth are said to be normally magnetized, whereas those that are relatively young and exhibit a magnetism that is 180° different from the current magnetic field are said to be reversely magnetized. Young, normally magnetized rocks in the Northern Hemisphere have a magnetic field that is oriented down and directed to the north. Young, reversely magnetized rocks in the Northern Hemisphere have a magnetic field that is oriented up and directed to the south. As mentioned earlier, studies of the age and polarity of volcanic rocks collected from all over the world indicate that the Earth's magnetic field reversed itself a number of times in the past. Nearly all samples of volcanic rocks younger than 690,000 years are normally magnetized, corresponding to the fact that the Earth's magnetic field has been normal for that time period. Volcanic rock samples 690,000 to 890,000 years old are reversely magnetized, reflective of the reversed orientation of the Earth's magnetic field during that interval.

Vine-Matthews hypothesis

In 1963 the British geophysicists Frederick J. Vine and Drummond H. Matthews proposed that the magnetic anomalies in oceanic regions were produced as a result of the extrusion of volcanic rocks on the ocean floor during times of alternating normal and reversed magnetic fields. In this model, the magnetism of lavas extruded when the magnetic field of the Earth is normal adds to the magnetism generated in the planet's liquid core, resulting in a positive magnetic anomaly. By contrast, the magnetism of lavas extruded when the Earth's field is reversed subtracts from the magnetism produced in the core, giving rise to a negative magnetic anomaly.

The Vine-Matthews hypothesis provides an explanation for the symmetry of the magnetic anomalies over mid-oceanic ridges. As magnetized lavas are erupted during the formation of new oceanic crust at the ridge crests, the resulting volcanic rocks are carried away from the crests in both directions. The pattern of the magnetic anomalies is virtually identical to the pattern of the geomagnetic time scale. These observations can only be explained if volcanic activity occurred principally at the crests of mid-oceanic ridges and if seafloor spreading occurred during alternating periods of normal and reversed polarity of the Earth. Moreover, magnetic profiles calculated by computers using assumed strips of alternating normally and reversely magnetized volcanic rocks matched observed magnetic profiles almost exactly, providing further support for the model set forth by Vine and Matthews. (C.K.S.)

BIBLIOGRAPHY

The figure of the Earth: Recent standard works include G. BOMFORD, *Geodesy*, 4th ed. (1980, reprinted 1983); WEIKKO A. HEISKANEN and HELMUT MORITZ, *Physical Geodesy* (1967); WOLFGANG TORGE, *Geodesy, an Introduction* (1980; originally published in German, 1975), a classical treatment; and PETR VANÍČEK and EDWARD J. KRAKIWSKY, *Geodesy, the Concepts*, rev. ed. (1986). See also INTERNATIONAL ASSOCIATION OF GEODESY, *Systeme geodesique de reference 1967: Geodetic Reference System 1967* (1971).

The Earth's gravitational field: Texts include MICHELE CAPUTO, *The Gravity Field of the Earth, from Classical and Modern Methods* (1967); W.A. HEISKANEN and F.A. VENING MEINESZ, *The Earth and Its Gravity Field* (1958); and W.M. TELFORD *et al.*, *Applied Geophysics* (1976). The *Journal of Geophysical Research* contains much recent work, including the following articles: LEROY M. DORMAN and BRIAN T.R. LEWIS, "Experimental Isostasy, 1: Theory of the Determination of the Earth's Isostatic Response to a Concentrated Load," 75(17):3357–65 (June 10, 1970), "Experimental Isostasy, 2: An Isostatic Model for the U.S.A. Derived from Gravity and Topographic Data," 75(17):3367–86 (June 10, 1970), and "Experimental Isostasy, 3: Inversion of the Isostatic Green Function and Lateral Density Changes," 77(17):3068–77 (June 10, 1972); G.D. KARNER and A.B. WATTS, "Gravity Anomalies and Flexure of the Lithosphere at Mountain Ranges," 88(B12):10,449–10,477 (Dec. 10, 1983); and MARCIA MCNUTT, "Implications of Regional Gravity for State of Stress in the Earth's Crust and Upper Mantle," 85(B11):6377–96 (Nov. 10, 1980).

(G.D.G.)

The Earth's magnetic field: Brief surveys are SYDNEY CHAPMAN, *The Earth's Magnetism*, 2nd ed. rev. (1951, reprinted 1961); and J.A. JACOBS, *The Earth's Core and Geomagnetism* (1963), and *Reversals of the Earth's Magnetic Field* (1984). For more in-depth coverage, consult SYDNEY CHAPMAN and JULIUS BARTELS, *Geomagnetism*, vol. 1, *Geomagnetic and Related Phenomena* (1940, reprinted 1962); GEORGE D. GARLAND, *Introduction to Geophysics: Mantle, Core and Crust*, 2nd ed. (1979), an informative overview; S. MATSUSHITA, "Solar Quiet and Lunar Daily Variation Fields," ch. III-1 in S. MATSUSHITA and WALLACE H. CAMPBELL, *Physics of Geomagnetic Phenomena*, vol. 1 (1967), pp. 301–424, an extensive treatment; RONALD T. MERRILL and MICHAEL W. MCELHINNY, *The Earth's Magnetic Field: Its History, Origin, and Planetary Perspective* (1983), an attempt to bridge the gap between dynamo theory and paleomagnetology; W.D. PARKINSON, *Introduction to Geomagnetism* (1983), a logical, systematic, and highly readable treatment; and S.K. RUNCORN, K.M. CREER, and J.A. JACOBS (eds.), *The Earth's Core: Its Structure, Evolution, and Magnetic Field* (1982), an excellent introduction to the modern literature on the topic. See also S.J. PEALE, "Consequences of Tidal Evolution," ch. 12 in MARGARET G. KIVELSON (ed.), *The Solar System: Observations and Interpretations* (1986), pp. 275–288; and K.A. WIENERT, *Notes on Geomagnetic Observatory and Survey Practice* (1970), an explanation of the basic principles of geomagnetic survey techniques.

(R.L.M.)

Structure and composition of the solid Earth: DAVID G. SMITH (ed.), *The Cambridge Encyclopedia of Earth Sciences* (1981), is a highly readable, illustrated collection of articles on the Earth's surface and interior. Two collections of articles from *Scientific American* are *The Dynamic Earth* (1983), on the geologic processes occurring inside and around the planet; and ROBERT DECKER and BARBARA DECKER (eds.), *Volcanoes and the Earth's Interior* (1982), on volcanoes and the rocks occurring in the mantle. MINORU OZIMA, *The Earth: Its Birth and Growth* (1981; originally published in Japanese, 1979), gives a general account of the evolution of the Earth based on geochemical research. MARTIN H.P. BOTT, *Interior of the Earth: Its Structure, Constitution, and Evolution*, 2nd ed. (1982), provides a thorough summary of the nature of the Earth's interior. Two works with a more technical slant are JEAN-CLAUDE DE BREMAECKER, *Geophysics, the Earth's Interior* (1985), an introductory textbook; and DONALD L. TURCOTTE and GERALD SCHUBERT, *Geodynamics: Applications of Continuum Physics to Geological Problems* (1982), a mathematical text on processes occurring within the Earth. See also G.C. BROWN and A.E. MUSSETT, *The Inaccessible Earth* (1981), on both geophysics and geochemistry.

(R.J.)

Major geologic features of the Earth's exterior: DAVID ALT, *Physical Geology: A Process Approach* (1982), is an up-to-date introduction to the principles of physical geology. DANIEL S. BARKER, *Igneous Rocks* (1983), discusses the characteristics of igneous rocks and their relationship to plate tectonics. MARLAND P. BILLINGS, *Structural Geology*, 3rd ed. (1972), is a classic introduction to structural geology, which describes the deformation of the Earth's crust. A. HALLAM, *A Revolution in the Earth Sciences: From Continental Drift to Plate Tectonics* (1973), is an excellent historical review of the development of ideas on continental drift, polar wandering, seafloor spreading, and plate tectonics. ARTHUR HOLMES, *Principles of Physical Geology*, 2nd rev. ed. (1965), is a beautifully written, though somewhat dated, introduction to physical geology. ROBLEY K. MATTHEWS, *Dynamic Stratigraphy: An Introduction to Sedimentation and Stratigraphy*, 2nd ed. (1984), relates the study of the sequence of layered rocks to plate tectonics. WILLIAM D. THORNBURY, *Principles of Geomorphology*, 2nd ed. (1969), is a classic account of the shaping of the Earth's surface by natural forces, including a description of the relationship between structural geology and landforms. TJEERD H. VAN ANDEL, *New Views on an Old Planet: Continental Drift and the History of Earth* (1985), is an excellent introduction to the ideas of continental drift, polar wandering, seafloor spreading, and plate tectonics from the viewpoint of an oceanographer. BRIAN F. WINDLEY, *The Evolving Continents*, 2nd ed. (1984), relates the geology of various parts of the Earth to plate tectonics and the Wilson cycle of ocean-basin formation. Studies of historical geology include DON L. EICHER, A. LEE MCALESTER, and MARCIA L. ROTTMAN, *The History of the Earth's Crust* (1984); and CARL K. SEYFERT and LESLIE A. SIRKIN, *Earth History and Plate Tectonics* (1979), which relates the history of the Earth to plate tectonics, especially to the Wilson cycle.

(C.K.S.)

The Earth Sciences

The broad aim of the Earth sciences is to understand the present features and the past evolution of the Earth and to use this knowledge, where appropriate, for the benefit of humankind. Thus the basic concerns of the Earth scientist are to observe, describe, and classify all the features of the Earth, whether characteristic or not, to generate hypotheses with which to explain their presence and their development, and to devise means of checking opposing ideas for their relative validity. In this way the most plausible, acceptable, and long-lasting ideas are developed.

The physical environment in which humans live includes not only the immediate surface of the solid Earth, but also the ground beneath it and the water and air above it. Early man was more involved with the practicalities of life than with theories, and thus his survival depended on his ability to obtain metals from the ground to produce, for example, alloys, such as bronze from copper and tin, for tools and armour, to find adequate water supplies for establishing dwelling sites, and to forecast the weather, which had a far greater bearing on human life in earlier times than it has today. Such situations represent the foundations of the three principal component disciplines of the modern Earth sciences—the geologic, hydrologic, and atmospheric—which form the tripartite framework of the present article.

The rapid development of science as a whole over the past century and a half has given rise to an immense number of specializations and subdisciplines, with the result that the modern Earth scientist, perhaps unfortunately, tends to know a great deal about a very small area of study but only a little about most other aspects of the entire field. It is therefore very important for the lay person and the researcher alike to be aware of the complex interlinking network of disciplines that make up the Earth sciences today, and that is the purpose of this article. Only when one is aware of the marvelous complexity of the Earth sciences and yet can understand the breakdown of the component disciplines is one in a position to select those parts of the subject that are of greatest personal interest.

It is worth emphasizing two important features that the three divisions of the Earth sciences have in common. First is the inaccessibility of many of the objects of study. Many rocks, as well as water and oil reservoirs, are at great depths in the Earth, while air masses circulate at vast heights above it. Thus the Earth scientist has to have a good three-dimensional perspective. Second, there is the fourth dimension: time. The Earth scientist is responsible for working out how the Earth evolved over millions of years. For example, what were the physical and chemical conditions operating on the Earth and the Moon 3,500,000,000 years ago? How did the oceans form, and how did their chemical composition change with time? How has the atmosphere developed? And finally, how did life on Earth begin, and from what did man evolve?

Today the Earth sciences are divided into many disciplines, which are themselves divisible into six groups:

(1) Those subjects that deal with the water and air at or above the solid surface of the Earth. These include the study of the water on and within the ground (hydrology), the glaciers and ice caps (glaciology), the oceans (oceanography), the atmosphere and its phenomena (meteorology), and the world's climates (climatology). In this article such fields of study are grouped under the hydrologic and atmospheric sciences and are treated separately from the geologic sciences, which focus on the solid Earth.

(2) Disciplines concerned with the physical-chemical makeup of the solid Earth, which include the study of minerals (mineralogy), the three main groups of rocks (igneous, sedimentary, and metamorphic petrology), the chemistry of rocks (geochemistry), the structures in rocks (structural geology), and the physical properties of rocks at the Earth's surface and in its interior (geophysics).

(3) The study of landforms (geomorphology), which is concerned with the description of the features of the present terrestrial surface and an analysis of the processes that gave rise to them.

(4) Disciplines concerned with the geologic history of the Earth, including the study of fossils and the fossil record (paleontology), the development of sedimentary strata deposited typically over millions of years (stratigraphy), and the isotopic chemistry and age dating of rocks (geochronology).

(5) Applied Earth sciences dealing with current practical applications beneficial to society. These include the study of fossil fuels (oil, natural gas, and coal); oil reservoirs; mineral deposits; geothermal energy for electricity and heating; the structure and composition of bedrock for the location of bridges, nuclear reactors, roads, dams, and skyscrapers and other buildings; hazards involving rock and mud avalanches, volcanic eruptions, earthquakes, and the collapse of tunnels; and coastal, cliff, and soil erosion.

(6) The study of the rock record on the Moon and the planets and their satellites (astrogeology). This field includes the investigation of relevant terrestrial features—namely, tektites (glassy objects resulting from meteorite impacts) and astroblemes (meteorite craters).

With such intergradational boundaries between the divisions of the Earth sciences (which, on a broader scale, also intergrade with physics, chemistry, biology, mathematics, and certain branches of engineering), researchers today must be versatile in their approach to problems. Hence, an important aspect of training within the Earth sciences is an appreciation of their multidisciplinary nature.

(B.F.W.)

This article is divided into the following sections:

History of the Earth sciences 616
- Origins in prehistoric times 616
- Antiquity 616
  - Geologic sciences
  - Hydrologic and atmospheric sciences
- The 16th–18th centuries 617
  - Geologic sciences
  - Hydrologic sciences
  - Atmospheric sciences
- The 19th century 619
  - Geologic sciences
  - Hydrologic sciences
  - Atmospheric sciences
- The 20th century: modern trends and developments 622
  - Geologic sciences
  - Hydrologic sciences
  - Atmospheric sciences

Geologic sciences 626
- Study of the composition of the Earth 627
  - Mineralogy
  - Petrology
  - Economic geology
  - Geochemistry
- Study of the structure of the Earth 631
  - Geodesy
  - Geophysics
  - Structural geology
  - Volcanology
- Study of surface features and processes 634
  - Geomorphology
  - Glacial geology
- Earth history 634
  - Historical geology and stratigraphy
  - Paleontology
  - Astrogeology

Practical applications 636
Exploration for energy and mineral sources
Earthquake prediction and control
Other areas of application
Hydrologic sciences 637
Study of the waters close to the land surface 638
Evaluation of the catchment water balance
Modeling catchment hydrology
Water quality
Study of lakes 641
The history of lakes
The physical characteristics of lakes
Water and energy fluxes in lakes
The water quality of lakes
Study of the oceans and seas 642
The origin of the ocean basins
The physical properties of seawater
The circulation of the oceans
Biogeochemical cycles in the oceans
Remote sensing of the oceans
Study of ice on the Earth's land surface 643
The accumulation of ice
The movement of glaciers
Practical applications 643
Development and management of water resources
Concern over groundwater quantity and quality
Studying the causes of droughts and other climatic patterns
Atmospheric sciences 644
Study of the evolution of the atmosphere 645
Study of surface budgets 645
Heating by radiation, conduction, and convection
Other factors affecting atmospheric processes and conditions
Study of the vertical structure of the atmosphere 646
The planetary boundary layer
The free atmosphere and the tropopause
The upper regions of the atmosphere
Study of the horizontal structure of the atmosphere 648
Asymmetrical distribution of solar heating and its effects
Atmospheric circulation patterns
Study of cloud processes 650
Elements of cloud formation
Precipitation processes in clouds
Procedures of cloud physics research
Study of climate and climatic change 651
Local determining factors
Climatic change and its causes
Practical applications 652
Weather forecasting
Weather modification
Bibliography 653

History of the Earth sciences

ORIGINS IN PREHISTORIC TIMES

The origins of the Earth sciences lie in the myths and legends of the distant past. The creation story, which can be traced to a Babylonian epic of the 22nd century BC and which is told in the first chapter of Genesis, has proved most influential. The story is cast in the form of Earth history and thus was readily accepted as an embodiment of scientific as well as of theological truth.

Earth scientists later made innumerable observations of natural phenomena and interpreted them in an increasingly multidisciplinary manner. The Earth sciences, however, were slow to develop largely because the progress of science was constrained by whatever society would tolerate or support at any one time.

ANTIQUITY

Classification of rocks and minerals by Theophrastus

Geologic sciences. *Knowledge of Earth composition and structure.* The oldest known treatise on rocks and minerals is the *De lapidibus* ("On Stones") of the Greek philosopher Theophrastus (*c.* 372–*c.* 287 BC). Written probably in the early years of the 3rd century, this work remained the best study of mineral substances for almost 2,000 years. Although reference is made to some 70 different materials, the work is more an effort at classification than systematic description.

In early Chinese writings on mineralogy, stones and rocks were distinguished from metals and alloys, and further distinctions were made on the basis of colour and other physical properties. The speculations of Cheng Ssu-hsiao (died AD 1332) on the origin of ore deposits were more advanced than those of his contemporaries in Europe. In brief, his theory was that ore is deposited from groundwater circulating in subsurface fissures.

Speculations on earthquakes and volcanic eruptions

Ancient accounts of earthquakes and volcanic eruptions are sometimes valuable as historical records but tell little about the causes of these events. Aristotle (384–322 BC) and Strabo (64 BC–*c.* AD 21) held that volcanic explosions and earthquakes alike are caused by spasmodic motions of hot winds that move underground and occasionally burst forth in volcanic activity attended by Earth tremors. Classical and medieval ideas on earthquakes and volcanoes were brought together in William Caxton's *Mirrour of the World* (1480). Earthquakes are here again related to movements of subterranean fluids. Streams of water in the Earth compress the air in hidden caverns. If the roofs of the caverns are weak, they rupture, causing cities and castles to fall into the chasms; if strong, they merely tremble and shake from the heaving by the wind below. Volcanic action follows if the outburst of wind and water from the depths is accompanied by fire and brimstone from hell.

Chinese invention of the seismograph

The Chinese have the distinction of keeping the most faithful records of earthquakes and of inventing the first instrument capable of detecting them. Records of the dates on which major quakes rocked China date to 780 BC. To detect quakes at a distance, the mathematician, astronomer, and geographer Chang Heng (AD 78–139) invented what has been called the first seismograph.

Knowledge of Earth history. The occurrence of seashells embedded in the hard rocks of high mountains aroused the curiosity of early naturalists and eventually set off a controversy on the origin of fossils that continued through the 17th century. Xenophanes of Colophon (flourished *c.* 560 BC) was credited by later writers with observing that seashells occur "in the midst of earth and in mountains." He is said to have believed that these relics originated during a catastrophic event that caused the Earth to be mixed with the sea and then to settle, burying organisms in the drying mud. For these views Xenophanes is sometimes called the father of paleontology.

Petrified wood was described by Chinese scholars as early as the 9th century AD, and around 1080 Shen Kua cited fossilized plants as evidence for change in climate. Other kinds of fossils that attracted the attention of early Chinese writers include spiriferoid brachiopods ("stone swallows"), cephalopods, crabs, and the bones and teeth of reptiles, birds, and mammals. Although these objects were commonly collected simply as curiosities or for medicinal purposes, Shen Kua recognized marine invertebrate fossils for what they are and for what they imply historically. Observing seashells in strata of the T'ai-hang Shan range, he concluded that this region, though now far from the sea, must once have been a shore.

Formation of river deltas and alluvial plains

Knowledge of landforms and of land–sea relations. Changes in the landscape and in the position of land and sea related to erosion and deposition by streams were recognized by some early writers. The Greek historian Herodotus (*c.* 484–*c.* 426 BC) correctly concluded that the northward bulge of Egypt into the Mediterranean is caused by the deposition of mud carried by the Nile.

The early Chinese writers were not outdone by the Romans and Greeks in their appreciation of changes wrought by erosion. In the *Chin shu* ("History of the Chin Dynasty"), it is said of Tu Yü (AD 222–284) that when he ordered monumental stelae to be carved with the records of his successes, he had one buried at the foot of a mountain and the other erected on top. He predicted that in time they would likely change their relative positions, because the high hills will become valleys and the deep valleys will become hills.

Aristotle guessed that changes in the position of land and sea might be cyclical in character, thus reflecting some sort of natural order. If the rivers of a moist region should

build deltas at their mouths, he reasoned, seawater would be displaced and the level of the sea would rise to cover some adjacent dry region. A reversal of climatic conditions might cause the sea to return to the area from which it had previously been displaced and retreat from the area previously inundated. The idea of a cyclical interchange between land and sea was elaborated in the *Discourses of the Brothers of Purity*, a classic Arabic work written between AD 941 and 982 by an anonymous group of scholars at Basra (Iraq). The rocks of the lands disintegrate and rivers carry their wastage to the sea, where waves and currents spread it over the seafloor. There the layers of sediment accumulate one above the other, harden, and, in the course of time, rise from the bottom of the sea to form new continents. Then the process of disintegration and leveling begins again.

Hydrologic and atmospheric sciences. The only substance known to the ancient philosophers in its solid, liquid, and gaseous states, water is prominently featured in early theories about the origin and operations of the Earth. Thales of Miletus (*c.* 624–*c.* 545 BC) is credited with a belief that water is the essential substance of the Earth, and Anaximander of Miletus (*c.* 610–545 BC) held that water was probably the source of life. In the system proposed by Empedocles of Agrigentum (*c.* 490–430 BC), water shared the primacy Thales had given it with three other elements: fire, air, and earth. The doctrine of the four earthly elements was later embodied in the universal system of Aristotle and thereby influenced Western scientific thought until late in the 17th century.

Knowledge of the hydrologic cycle. The idea that the waters of the Earth undergo cyclical motions, changing from seawater to vapour to precipitation and then flowing back to the ocean, is probably older than any of the surviving texts that hint at or frame it explicitly.

The idea of the hydrological cycle developed independently in China as early as the 4th century BC and was explicitly stated in the *Lü-shih Ch'un Ch'iu* ("The Spring and Autumn Annals of Mr. Lü"), written in the 3rd century BC. A circulatory system of a different kind, involving movements of water on a large scale within the Earth, was envisioned by Plato (*c.* 428–348/347 BC). In one of his two explanations for the origin of rivers and springs, he described the Earth as perforated by passages connecting with Tartarus, a vast subterranean reservoir.

Origin of rivers, springs, and precipitation

A coherent theory of precipitation is found in the writings of Aristotle. Moisture on the Earth is changed to airy vapour by heat from above. Because it is the nature of heat to rise, the heat in the vapour carries it aloft. When the heat begins to leave the vapour, the vapour turns to water. The formation of water from air produces clouds. Heat remaining in the clouds is further opposed by the cold inherent in the water and is driven away. The cold presses the particles of the cloud closer together, restoring in them the true nature of the element water. Water naturally moves downward, and so it falls from the cloud as raindrops. Snow falls from clouds that have frozen.

In Aristotle's system the four earthly elements were not stable but could change into one another. If air can change to water in the sky, it should also be able to change into water underground.

The origin of the Nile. Of all the rivers known to the ancients, the Nile was most puzzling with regard to its sources of water. Not only does this river maintain its course up the length of Egypt through a virtually rainless desert, but it rises regularly in flood once each year.

Speculations on the strange behaviour of the Nile were many, varied, and mostly wrong. Thales suggested that the strong winds that blow southward over the delta in summertime hold back the flow of the river and cause the waters to rise upstream in flood. Oenopides of Chios (flourished *c.* 475 BC) thought that heat stored in the ground during the winter dries up the underground veins of water so that the river shrinks. In the summer the heat disappears, and water flows up into the river, causing floods. In the view of Diogenes of Apollonia (flourished *c.* 435 BC), the Sun controls the regimen of the stream. The idea that the Nile waters connect with the sea is an old one, tracing back to the geographic concepts of Hecataeus of Miletus (*c.* 520 BC). Reasonable explanations related the discharge of the Nile to precipitation in the headwater regions, as snow (Anaxagoras of Clazomenae, *c.* 500–428 BC) or from rain that filled lakes supposed to have fed the river (Democritus of Abdera, *c.* 460–*c.* 357 BC). Eratosthenes (*c.* 276–194 BC), who had prepared a map of the Nile Valley southward to the latitude of modern Khartoum, anticipated the correct answer when he reported that heavy rains had been observed to fall in the upper reaches of the river and that these were sufficient to account for the flooding.

Knowledge of the tides. The tides of the Mediterranean, being inconspicuous in most places, attracted little notice from Greek and Roman naturalists. Poseidonius (135–50 BC) first correlated variations in the tides with phases of the Moon.

By contrast, the tides along the eastern shores of Asia generally have a considerable range and were the subject of close observation and much speculation among the Chinese. In particular, the tidal bore on the Ch'ien-t'ang Chiang (Ch'ien-t'ang River) near Hang-chou attracted early attention; with its front ranging up to 3.7 metres in height, this bore is one of the largest in the world. As early as the 2nd century BC, the Chinese had recognized a connection between tides and tidal bores and the lunar cycle.

Prospecting for groundwater. Although the origin of water in the Earth that seeps or springs from the ground was long the subject of much fanciful speculation, the arts of finding and managing groundwater were already highly developed in the 8th century BC. The construction of long, hand-dug underground aqueducts (*qanats*) in Armenia and Persia represents one of the great hydrologic achievements of the ancient world. After some 3,000 years *qanats* are still a major source of water in Iran.

Construction of *qanats*

In the 1st century BC, Vitruvius (Marcus Vitruvius Pollio), a Roman architect and engineer, described methods of prospecting for groundwater in his *De architectura libri decem* (*The Architecture of Marcus Vitruvius Pollio, in Ten Books*). To locate places where wells should be dug, he recommended looking for spots where mist rises in early morning. More significantly, Vitruvius had learned to associate different quantities and qualities of groundwater with different kinds of rocks and topographic situations.

After the inspired beginnings of the ancient Greeks, Romans, Chinese, and Arabs, little or no new information was collected, and no new ideas were produced throughout the Middle Ages, appropriately called the Dark Ages. It was not until the Renaissance in the early 16th century that the Earth sciences began to develop again.

THE 16TH–18TH CENTURIES

Geologic sciences. *Ore deposits and mineralogy.* A common belief among alchemists of the 16th and 17th centuries held that metalliferous deposits were generated by heat emanating from the centre of the Earth but activated by the heavenly bodies.

The German scientist Georgius Agricola has with much justification been called the father of mineralogy. Of his seven geologic books, *De natura fossilium* (1546; "On Natural Fossils") contains his major contributions to mineralogy and, in fact, has been called the first textbook on that subject. In Agricola's time and well into the 19th century, "fossil" was a term that could be applied to any object dug from the Earth. Thus Agricola's classification of fossils provided pigeonholes for organic remains, such as ammonites, and for rocks of various kinds in addition to minerals. Individual kinds of minerals, their associations and manners of occurrence, are described in detail, many for the first time.

Agricola's *De natura fossilium*

With the birth of analytical chemistry toward the latter part of the 18th century, the classification of minerals on the basis of their composition at last became possible. The German geologist Abraham Gottlob Werner was one of those who favoured a chemical classification in preference to a "natural history" classification based on external appearances. His list of several classifications, published posthumously, recognized 317 different substances ordered in four classes.

Paleontology and stratigraphy. During the 17th century

the guiding principles of paleontology and historical geology began to emerge in the work of a few individuals. Nicolaus Steno, a Danish scientist and theologian, presented carefully reasoned arguments favouring the organic origin of what are now called fossils. Also, he elucidated three principles that made possible the reconstruction of certain kinds of geologic events in a chronological order. In his *Canis carcariae dissectum caput* (1667; "Dissected Head of a Dog Shark"), he concluded that large tongue-shaped objects found in the strata of Malta were the teeth of sharks, whose remains were buried beneath the seafloor and later raised out of the water to their present sites. This excursion into paleontology led Steno to confront a broader question. How can one solid body, such as a shark's tooth, become embedded in another solid body, such as a layer of rock? He published his answers in 1669 in a paper entitled "De solido intra naturaliter contento dissertationis" ("A Preliminary Discourse Concerning a Solid Body Enclosed by Processes of Nature Within a Solid"). Steno cited evidence to show that when the hard parts of an organism are covered with sediment, it is they and not the aggregates of sediment that are firm. Consolidation of the sediment into rock may come later, and, if so, the original solid fossil becomes encased in solid rock. He recognized that sediments settle from fluids layer by layer to form strata that are originally continuous and nearly horizontal. His principle of superposition of strata states that in a sequence of strata, as originally laid down, any stratum is younger than the one on which it rests and older than the one that rests upon it.

Steno's principles of superposition, continuity, and horizontality of strata

In 1667 and 1668 the English physicist Robert Hooke read papers before the Royal Society in which he expressed many of the ideas contained in Steno's works. Hooke argued for the organic nature of fossils. Elevation of beds containing marine fossils to mountainous heights he attributed to the work of earthquakes. Streams attacking these elevated tracts wear down the hills, fill depressions with sediment, and thus level out irregularities of the landscape.

Earth history according to Werner and James Hutton. The two major theories of the 18th century were the Neptunian and the Plutonian. The Neptunists, led by Werner and his students, maintained that the Earth was originally covered by a turbid ocean. The first sediments deposited over the irregular floor of this universal ocean formed the granite and other crystalline rocks. Then as the ocean began to subside, "Stratified" rocks were laid down in succession. The "Volcanic" rocks were the youngest; Neptunists took small account of volcanism and thought that lava was formed by the burning of coal deposits underground.

Neptunists and the universal ocean

The Scottish scientist James Hutton, leader of the Plutonists, viewed the Earth as a dynamic body that functions as a heat machine. Streams wear down the continents and deposit their waste in the sea. Subterranean heat causes the outer part of the Earth to expand in places, uplifting the compacted marine sediments to form new continents. Hutton recognized that granite is an intrusive igneous rock and not a primitive sediment as the Neptunists claimed. Intrusive sills and dikes of igneous rock provide evidence for the driving force of subterranean heat. Hutton viewed great angular unconformities separating sedimentary sequences as evidence for past cycles of sedimentation, uplift, and erosion. His *Theory of the Earth,* published as an essay in 1788, was expanded to a two-volume work in 1795. John Playfair, a professor of natural philosophy, defended Hutton against the counterattacks of the Neptunists, and his *Illustrations of the Huttonian Theory* (1802) is the clearest contemporary account of Plutonist theory.

Hydrologic sciences. The idea that there is a circulatory system within the Earth, by which seawater is conveyed to mountaintops and there discharged, persisted until early in the 18th century. Two questions left unresolved by this theory were acknowledged even by its advocates. How is seawater forced uphill? How is the salt lost in the process?

Ideas on the origin of springs

The rise of subterranean water. René Descartes supposed that the seawater diffused through subterranean channels into large caverns below the tops of mountains. The Jesuit philosopher Athanasius Kircher in his *Mundus subterraneus* (1664; "Subterranean World") suggested that the tides pump seawater through hidden channels to points of outlet at springs. To explain the rise of subterranean water beneath mountains, the chemist Robert Plot appealed to the pressure of air, which forces water up the insides of mountains. The idea of a great subterranean sea connecting with the ocean and supplying it with water together with all springs and rivers was resurrected in 1695 in John Woodward's *Essay towards a Natural History of the Earth and Terrestrial Bodies.*

The French Huguenot Bernard Palissy maintained, to the contrary, that rainfall is the sole source of rivers and springs. In his *Discours admirables* (1580; *Admirable Discourses*) he described how rainwater falling on mountains enters cracks in the ground and flows down along these until, diverted by some obstruction, it flows out on the surface as springs. Palissy scorned the idea that seawater courses in veins to the tops of mountains. For this to be true, sea level would have to be higher than mountaintops—an impossibility. In his *Discours* Palissy suggested that water would rise above the level at which it was first encountered in a well provided the source of the groundwater came from a place higher than the bottom of the well. This is an early reference to conditions essential to the occurrence of artesian water, a popular subject among Italian hydrologists of the 17th and 18th centuries.

Explanation of artesian flow

In the latter part of the 17th century, Pierre Perrault and Edmé Mariotte conducted hydrologic investigations in the basin of the Seine River that established that the local annual precipitation was more than ample to account for the annual runoff.

Evaporation from the sea. The question remained as to whether the amount of water evaporated from the sea is sufficient to account for the precipitation that feeds the streams. The English astronomer-mathematician Edmond Halley measured the rate of evaporation from pans of water exposed to the air during hot summer days. Assuming that this same rate would obtain for the Mediterranean, Halley calculated that some 5,280,000,000 tons of water are evaporated from this sea during a summer day. Assuming further that each of the nine major rivers flowing into the Mediterranean has a daily discharge 10 times that of the Thames, he calculated that a daily inflow of fresh water back into that sea would be 1,827,000,000 tons, only slightly more than a third of the amount lost by evaporation. Halley went on to explain what happens to the remainder. A part falls back into the sea as rain before it reaches land. Another part is taken up by plants.

Halley's explanation of ocean salinity

In the course of the hydrologic cycle, Halley reasoned, the rivers constantly bring salt into the sea in solution, but the salt is left behind when seawater evaporates to replenish the streams with rainwater. Thus the sea must be growing steadily saltier.

Atmospheric sciences. *Water vapour in the atmosphere.* After 1760 the analytical chemists at last demonstrated that water and air are not the same substance in different guises. Long before this development, however, investigators had begun to draw a distinction between water vapour and air. Otto von Guericke, a German physicist and engineer, produced artificial clouds by releasing air from one flask into another one from which the air had been evacuated. A fog then formed in the unevacuated flask. Guericke concluded that air cannot be turned into water, though moisture can enter the air and later be condensed into water. Guericke's experiments, however, did not answer the question as to how water enters the atmosphere as vapour. In "Les Météores" ("Meteorology," an essay published in the book *Discours de la methode* in 1637), Descartes envisioned water as composed of minute particles that were elongate, smooth, and separated by a highly rarified "subtle matter."

The same uncertainty as to how water gets into the air surrounded the question as to how it remains suspended as clouds. A popular view in the 18th century was that clouds are made of countless tiny bubbles that float in air. Guericke had suggested that the fine particles in his artificial clouds were bubbles. Other observers professed to have seen bubble-shaped particles of water vapour rising from warm water or hot coffee.

Cloud formation and motion

Pressure, temperature, and atmospheric circulation. If clouds are essentially multicompartmented balloons, their motions could be explained by the movements of winds blowing on them. Descartes suggested that the winds might blow upward as well as laterally, causing the clouds to rise or at least preventing them from descending. In 1749 Benjamin Franklin explained updrafts of air as due to local heating of the atmosphere by the Sun. Sixteen years later the Swiss-German mathematical physicist Johann Heinrich Lambert described the conditions necessary for the initiation of convection currents in the atmosphere. He reasoned that rising warm air flows into bordering areas of cooler air, increasing their downward pressure and causing their lower layers to flow into ascending currents, thus producing circulation.

Importance of the barometer

The fact that Lambert could appeal to changes in air pressure to explain circulation reflects an important change from the view still current in the late 16th century that air is weightless. This misconception was corrected after 1643 with the invention of the mercury barometer. It was soon discovered that the height of the barometer varied with the weather, usually standing at its highest during clear weather and falling to the lowest on rainy days.

Toward the end of the 18th century it was beginning to be understood that variations in the barometer must be related to the general motion and circulation of the atmosphere. That these variations could not be due solely to changes in humidity was the conclusion of the Swiss scientist Horace Bénédict de Saussure in his *Essais sur l'hygrométrie* (1783; "Essay on Hygrometry"). From experiments with changes of water vapour and pressure in air enclosed in a glass globe, Saussure concluded that changes in temperature must be immediately responsible for variations of the barometer and that these in turn must be related to the movement of air from one place to another.

THE 19TH CENTURY

Geologic sciences. *Crystallography and the classification of minerals and rocks.* The French scientist René-Just Haüy, whose treatises on mineralogy and crystallography appeared in 1801 and 1822, respectively, has been credited with advancing mineralogy to the status of a science and with establishing the science of crystallography. From his studies of the geometric relationships between planes of cleavage, he concluded that the ultimate particles forming a given species of mineral have the same shape and that variations in crystal habit reflect differences in the ways identical molecules are put together. In 1814 Jöns Jacob Berzelius of Sweden published a system of mineralogy offering a comprehensive classification of minerals based on their chemistry. Berzelius recognized silica as an acid and introduced into mineralogy the group known as silicates. At mid-century the American geologist James Dwight Dana's *System of Mineralogy,* in its third edition, was reorganized around a chemical classification, which thereafter became standard for handbooks.

Optical studies of rocks

The development of the polarizing microscope and the technique for grinding sections of rocks so thin as to be virtually transparent came in 1827 from studies of fossilized wood by William Nicol. In 1849 Clifton Sorby showed that minerals viewed in thin section could be identified by their optical properties, and soon afterward improved classifications of rocks were made on the basis of their mineralogic composition. The German geologist Ferdinand Zirkel's *Mikroscopische Beschaffenheit der Mineralien und Gesteine* (1873; "The Microscopic Nature of Minerals and Rocks") contains one of the first mineralogic classifications of rocks and marks the emergence of microscopic petrography as an established branch of science.

William Smith and faunal succession. In 1683 the zoologist Martin Lister proposed to the Royal Society that a new sort of map be drawn showing the areal distribution of the different kinds of British "soiles" (vegetable soils and underlying bedrock). The work proposed by Lister was not accomplished until 132 years later, when William Smith published his *Geologic Map of England and Wales with Part of Scotland* (1815). A self-educated surveyor and engineer, Smith had the habit of collecting fossils and making careful note of the strata that contained them. He discovered that the different stratified formations in England contain distinctive assemblages of fossils. His map, reproduced on a scale of five miles to the inch, showed 20 different rock units, to which Smith applied local names in common use—*e.g.,* London Clay and Purbeck Beds. In 1816 Smith published a companion work, *Strata Identified by Organized Fossils,* in which the organic remains characteristic of each of his rock units were illustrated. His generalization that each formation is "possessed of properties peculiar to itself [and] has the same organized fossils throughout its course" is the first clear statement of the principle of faunal sequence, which is the basis for worldwide correlation of fossiliferous strata into a coherent system. Smith thus demonstrated two kinds of order in nature: order in the spatial arrangement of rock units and order in the succession of ancient forms of life.

Smith's principle of faunal sequence was another way of saying that there are discontinuities in the sequences of fossilized plants and animals. These discontinuities were interpreted in two ways: as indicators of episodic destruction of life or as evidence for the incompleteness of the fossil record. Baron Georges Cuvier of France was one of the more distinguished members of a large group of naturalists who believed that paleontological discontinuities bore witness to sudden and widespread catastrophes. Cuvier's skill at comparative anatomy enabled him to reconstruct from fragmentary remains the skeletons of large vertebrate animals found at different levels in the Cenozoic sequence of northern France. From these studies he discovered that the fossils in all but the youngest deposits belong to species now extinct. Moreover, these extinct species have definite ranges up and down in the stratigraphic column. Cuvier inferred that the successive extinctions were the result of convulsions that caused the strata of the continents to be dislocated and folded and the seas to sweep across the continents and just as suddenly subside.

Cuvier's catastrophic viewpoint

Charles Lyell and uniformitarianism. In opposition to the catastrophist school of thought, the British geologist Charles Lyell proposed a uniformitarian interpretation of geologic history in his *Principles of Geology* (3 vols., 1830–33). His system was based on two propositions: the causes of geologic change operating include all the causes that have acted from the earliest time; and these causes have always operated at the same average levels of energy. These two propositions add up to a "steady-state" theory of the Earth. Changes in climate have fluctuated around a mean, reflecting changes in the position of land and sea. Progress through time in the organic world is likewise an illusion, the effect of an imperfect paleontological record. The main part of the *Principles* was devoted less to theory than to procedures for inferring events from rocks; and for Lyell's clear exposition of methodology his work was highly regarded throughout its many editions, long after the author himself had abandoned antiprogressivist views on the development of life.

Louis Agassiz and the ice age. Huge boulders of granite resting upon limestone of the Jura Mountains were subjects of controversy during the 18th and early 19th centuries. Saussure described these in 1779 and called them erratics. He concluded that they had been swept to their present positions by torrents of water. Saussure's interpretation was in accord with the tenets of diluvial geologists, who interpreted erratics and sheets of unstratified sediment (till or drift) spread over the northern parts of Europe and North America as the work of the "Deluge."

Argument for large-scale glaciation of Europe and Asia

In 1837 the Swiss zoologist and paleontologist Louis Agassiz delivered a startling address before the Helvetian Society, proposing that, during a geologically recent stage of refrigeration, glacial ice had covered Eurasia from the North Pole to the shores of the Mediterranean and Caspian. Wherever erratics, till, and striated pavements of rock occur, sure evidence of this recent catastrophe exists. The reception accorded this address was glacial, too, and Alexander von Humboldt advised Agassiz to return to his fossil fishes. Instead, he began intensive field studies and in 1840 published his *Études sur les glaciers* ("Studies of Glaciers"), demonstrating that Alpine glaciers had been far more extensive in the past. That same year he visited the British Isles in the company of Buckland and extended

the glacial doctrine to Scotland, northern England, and Ireland. In 1846 he carried his campaign to North America and there found additional evidence for an ice age.

Geologic time and the age of the Earth. By mid-century the fossiliferous strata of Europe had been grouped into systems arrayed in chronological order. The stratigraphic column, a composite of these systems, was pieced together from exposures in different regions by application of the principles of superposition and faunal sequence. Time elapsed during the formation of a system became known as a period, and the periods were grouped into eras: the Paleozoic (Cambrian through Permian periods), Mesozoic (Triassic, Jurassic, and Cretaceous periods), and Cenozoic (Tertiary and Quaternary periods).

Charles Darwin's *Origin of Species* (1859) offered a theoretical explanation for the empirical principle of faunal sequence. The fossils of the successive systems are different not only because parts of the stratigraphic record are missing but also because most species have lost in their struggles for survival and also because those that do survive evolve into new forms over time. Darwin borrowed two ideas from Lyell and the uniformitarians: the idea that geologic time is virtually without limit and the idea that a sequence of minute changes integrated over long periods of time produce remarkable changes in natural entities.

Limitations based on the Earth's heat production

The evolutionists and the historical geologists were embarrassed when, beginning in 1864, William Thomson (later Lord Kelvin) attacked the steady-state theory of the Earth and placed numerical strictures on the length of geologic time. The Earth might function as a heat machine, but it could not also be a perpetual motion machine. Assuming that the Earth was originally molten, Thomson calculated that not less than 20,000,000 and not more than 400,000,000 years could have passed since the Earth first became a solid body. Other physicists of note put even narrower limits on the Earth's age ranging down to 15,000,000 or 20,000,000 years. All these calculations, however, were based on the common assumption, not always explicitly stated, that the Earth's substance is inert and hence incapable of generating new heat. Shortly before the end of the century this assumption was negated by the discovery of radioactive elements that disintegrate spontaneously and release heat to the Earth in the process.

Concepts of landform evolution. The scientific exploration of the American West following the end of the Civil War yielded much new information on the sculpture of the landscape by streams. John Wesley Powell in his reports on the Colorado River and Uinta Mountains (1875, 1876) explained how streams may come to flow across mountain ranges rather than detour around them. The Green River does not follow some structural crack in its gorge across the Uinta Mountains; instead it has cut its canyon as the mountain range was slowly bowed up. Given enough time, streams will erode their drainage basins to plains approaching sea level as a base. Grove Karl Gilbert's *Report on the Geology of the Henry Mountains* (1877) offered a detailed analysis of fluvial processes. According to Gilbert all streams work toward a graded condition, a state of dynamic equilibrium that is attained when the net effect of the flowing water is neither erosion of the bed nor deposition of sediment, when the landscape reflects a balance between the resistance of the rocks to erosion and the processes that are operative upon them. After 1884 William Morris Davis developed the concept of the geographical cycle, during which elevated regions pass through successive stages of dissection and denudation characterized as youthful, mature, and old. Youthful landscapes have broad divides and narrow valleys. With further denudation the original surface on which the streams began their work is reduced to ridgetops. Finally in the stage of old age, the region is reduced to a nearly featureless plain near sea level or its inland projection. Uplift of the region in any stage of this evolution will activate a new cycle. Davis' views dominated geomorphic thought until well into the 20th century, when quantitative approaches resulted in the rediscovery of Gilbert's ideas.

Gilbert's statement of dynamic equilibrium

Gravity, isostasy, and the Earth's figure. Discoveries of regional anomalies in the Earth's gravity led to the realization that high mountain ranges have underlying deficiencies in mass about equal to the apparent surface loads represented by the mountains themselves. In the 18th century the French scientist Pierre Bouguer had observed that the deflections of the pendulum in Peru are much less than they should be if the Andes represent a load perched on top of the Earth's crust. Similar anomalies were later found to obtain along the Himalayan front. To explain these anomalies it was necessary to assume that beneath some depth within the Earth pressures are hydrostatic (equal on all sides). If excess loads are placed upon the crust, as by addition of a continental ice cap, the crust will sink to compensate for the additional mass and will rise again when the load is removed. The tendency toward general equilibrium maintained through vertical movements of the Earth's outer layers was called isostasy in 1899 by Clarence Edward Dutton of the United States.

Concept of geosyncline and mountain building

Evidence for substantial vertical movements of the crust was supplied by studies of regional stratigraphy. In 1883 another American geologist, James Hall, had demonstrated that Paleozoic rocks of the folded Appalachians were several times as thick as sequences of the same age in the plateaus and plains to the west. It was his conclusion that the folded strata in the mountains must have accumulated in a linear submarine trough that filled with sediment as it subsided. Downward crustal flexures of this magnitude came to be called geosynclines.

Hydrologic sciences. *Darcy's law.* Quantitative studies of the movement of water in streams and aquifers led to the formulation of mathematical statements relating discharge to other factors. Henri-Philibert-Gaspard Darcy, a French hydraulic engineer, was the first to state clearly a law describing the flow of groundwater. Darcy's experiments, reported in 1856, were based on the ideas that an aquifer is analogous to a main line connecting two reservoirs at different levels and that an artesian well is like a pipe drawing water from a main line under pressure. His investigations of flow through stratified beds of sand led him to conclude that the rate of flow is directly proportional to the energy loss and inversely proportional to the length of the path of flow. Another French engineer, Arsène-Jules-Étienne-Juvénal Dupuit, extended Darcy's work and developed equations for underground flow toward a well, for the recharge of aquifers, and for the discharge of artesian wells. Philip Forchheimer, an Austrian hydrologist, introduced the theory of functions of a complex variable to analyze the flow by gravity of underground water toward wells and developed equations for determining the critical distance between a river and a well beyond which water from the river will not move into the well.

Surface water discharge. A complicated empirical formula for the discharge of streams resulted from the studies of Andrew Atkinson Humphreys and Henry Larcom Abbot in the course of the Mississippi Delta Survey of 1851–60. Their formula contained no term for roughness of channel and on this and other grounds was later found to be inapplicable to the rapidly flowing streams of mountainous regions. In 1869 Emile-Oscar Ganguillet and Rudolph Kutter developed a more generally applicable discharge equation following their studies of flow in Swiss mountain streams. Toward the end of the century, systematic studies of the discharge of streams had become common. In the United States the Geological Survey, following its establishment in 1879, became the principal agency for collecting and publishing data on discharge, and by 1906 stream gauging had become nationwide.

Foundations of oceanography. In 1807 Thomas Jefferson ordered the establishment of the U.S. Coast Survey (later Coast and Geodetic Survey and now the National Ocean Survey). Modeled after British and French agencies that had grown up in the 1700s, the agency was charged with the responsibilities of hydrographic and geodetic surveying, studies of tides, and preparation of charts. Beginning in 1842 the U.S. Navy undertook expansive oceanographic operations through its office of charts and instruments. Lieut. Matthew Fontaine Maury promoted international cooperation in gathering meteorologic and hydrologic data at sea. In 1847 Maury compiled the first wind and current charts for the North Atlantic and in 1854 issued the first depth map to 4,000 fathoms (7,300

Maury's compilation of oceanographic data

metres). His *Physical Geography of the Sea* (1855) is generally considered the first oceanographic textbook.

The voyage of the *Beagle* (1831–36) is remembered for Darwin's biological and geologic contributions. From his observations in the South Pacific, Darwin formulated a theory for the origin of coral reefs, which with minor changes has stood the test of time. He viewed the fringing reefs, barrier reefs, and atolls as successive stages in a developmental sequence. The volcanic islands around which the reef-building organisms are attached slowly sink, but at the same time the shallow-water organisms that form the reefs build their colonies upward so as to remain in the sunlit layers of water. With submergence of the island, what began as a fringing reef girdling a landmass at last becomes an atoll enclosing a lagoon.

Laying telegraphic cables across the Atlantic called for investigations of the configuration of the ocean floor, of the currents that sweep the bottom, and of the benthonic animals that might damage the cables. The explorations of the British ships *Lightning* and *Porcupine* in 1868 and 1869 turned up surprising oceanographic information. Following closely upon these voyages, the *Challenger* was authorized to determine "the conditions of the Deep Sea throughout the Great Ocean Basins."

The *Challenger* expedition (1872–76)

The *Challenger* left port in December of 1872 and returned in May 1876, after logging 127,600 kilometres (68,890 nautical miles). Under the direction of Wyville Thomson, Scottish professor of natural history, it occupied 350 stations scattered over all oceans except the Arctic. The work involved in analyzing the information gathered during the expedition was completed by Thomson's shipmate Sir John Murray, and the results filled 50 large volumes. Hundreds of new species of marine organisms were described, including new forms of life from deep waters. The temperature of water at the bottom of the oceans was found to be nearly constant below the 2,000-fathom level, averaging about 2.5° C (36.5° F) in the North Atlantic and 2° C (35° F) in the North Pacific. Soundings showed wide variations in depths of water, and from the dredgings of the bottom came new types of sediment—red clay as well as oozes made predominantly of the minute skeletons of foraminifera, radiolarians, or diatoms. Improved charts of the principal surface currents were produced, and the precise location of many oceanic islands was determined for the first time. Seventy-seven samples of seawater were taken at different stations from depths ranging downward to about 1.5 kilometres. The German-born chemist Wilhelm Dittmar conducted quantitative determinations of the seven major constituents (other than the hydrogen and oxygen of the water itself)—namely, sodium, calcium, magnesium, potassium, chloride, bromide, and sulfate. Surprisingly, the percentages of these components turned out to be nearly the same in all samples.

Tidal analysis and prediction

Efforts to analyze the rise and fall of the tides in mathematical terms reflecting the relative and constantly changing positions of Earth, Moon, and Sun, and thus to predict the tides at particular localities, has never been entirely successful because of local variations in configuration of shore and seafloor. Nevertheless, harmonic tidal analysis gives essential first approximations that are essential to tidal prediction. In 1884 a mechanical analog tidal prediction device was invented by William Ferrel of the U.S. Coast and Geodetic Survey, and improved models were used until 1965, when the work of the analog machines was taken over by electronic computers.

Atmospheric sciences. *Composition of the atmosphere.* Studies of barometric pressure by the British chemist and physicist John Dalton led him to conclude that evaporation and condensation of vapour do not involve chemical transformations. The introduction of vapour into the air by evaporation must change the average specific gravity of the air column and, without altering the height of that column, will change the reading of the barometer. In 1857 Rudolf Clausius, a German physicist, clarified the mechanics of evaporation in his kinetic theory of gases. Evaporation occurs when more molecules of a liquid are leaving its surface than returning to it, and the higher the temperature the more of these escaped molecules will be in space at any one time.

Sampling the atmosphere by balloon

Following the invention of the hot-air balloon by the Montgolfier brothers in 1783, balloonists produced some useful information on the composition and movements of the atmosphere. In 1804 the celebrated French chemist Joseph-Louis Gay-Lussac ascended to about 7,000 metres, took samples of air, and later determined that the rarefied air at that altitude contained the same percentage of oxygen (21.49 percent) as the air on the ground. Austrian meteorologist Julius von Hann, working with data from balloon ascents and climbing in the Alps and Himalayas, concluded in 1874 that about 90 percent of all the water vapour in the atmosphere is concentrated below 6,000 metres—from which it follows that high mountains can be barriers against the transport of water vapour.

Understanding of clouds, fog, and dew. Most of the names given to clouds (cirrus, cumulus, stratus, nimbus, and their combinations) were coined in 1803 by the English meteorologist Luke Howard. Howard's effort was not simply taxonomic; he recognized that clouds reflect in their shapes and changing forms "the general causes which effect all the variations of the atmosphere."

After Guericke's experiments it was widely believed that water vapour condenses into cloud as soon as the air containing it cools to the dew point. That this is not necessarily so was proved by Paul-Jean Coulier of France from experiments reported in 1875. Coulier found that the sudden expansion of air in glass flasks failed to produce an artificial cloud if the air in the system was filtered through cotton wool. He concluded that dust in the air was essential to the formation of cloud in the flask.

Discovery of condensation nuclei

From about the mid-1820s, efforts were made to classify precipitation in terms of the causes behind the lowering of temperature. In 1841 the American astronomer-meteorologist Elias Loomis recognized the following causes: warm air coming into contact with cold earth or water, responsible for fog; mixing of warm and cold currents, which commonly results in light rains; and sudden transport of air into high regions, as by flow up a mountain slope or by warm currents riding over an opposing current of cold air, which may produce heavy rains.

Recognition of the revolving nature of storms

Observation and study of storms. Storms, particularly tropical revolving storms, were subjects of much interest. As early as 1697 some of the more spectacular features of revolving storms were recorded in William Dampier's *New Voyage Round the World.* On July 4, 1687, Dampier's ship survived the passage of what he called a "tuffoon" off the coast of China. The captain's vivid account of this experience clearly describes the calm central eye of the storm and the passage of winds from opposite directions as the storm moved past. In 1828 Heinrich Wilhelm Dove, a Prussian meteorologist, recognized that tropical revolving storms are traveling systems with strong winds moving counterclockwise in the Northern Hemisphere and clockwise in the Southern Hemisphere. The whirlwind character of these storms was independently established by the American meteorologist William C. Redfield in the case of the September hurricane that struck New England in 1821. He noted that in central Connecticut the trees had been toppled toward the northwest, whereas some 80 kilometres westward they had fallen in the opposite direction. Redfield identified the belt between the Equator and the tropics as the region in which hurricanes are generated, and he recognized how the tracks of these storms tend to veer eastward when they enter the belt of westerly winds at about latitude 30° N. In 1849 Sir William Reid, a British meteorologist and military engineer, studied the revolving storms that occur south of the Equator in the Indian Ocean and confirmed that they have reversed rotations and curvatures of path compared with those of the Northern Hemisphere. Capt. Henry Piddington subsequently investigated revolving storms affecting the Bay of Bengal and Arabian Sea, and in 1855 he named these cyclones in his *Sailor's Horn-book for the Laws of Storms in all Parts of the World.*

Consideration of energy sources

Beginning in 1835 James Pollard Espy, an American meteorologist, began extensive studies of storms from which he developed a theory to explain their sources of energy. Radially convergent winds, he believed, cause the air to rise in their area of collision. Upward movement of moist

air is attended by condensation and precipitation. Latent heat released through the change of vapour to cloud or water causes further expansion and rising of the air. The higher the moist air rises the more the equilibrium of the system is disturbed, and this equilibrium cannot be restored until moist air at the surface ceases to flow toward the ascending column.

That radially convergent winds are not necessary to the rising of large air masses was demonstrated by Loomis in the case of a great storm that passed across the northeastern United States in December 1836. From his studies of wind patterns, changes of temperature, and changes in barometric pressure, he concluded that a cold northwest wind had displaced a wind blowing from the southeast by flowing under it. The southeast wind made its escape by ascending from the Earth's surface. Loomis had recognized what today would be called a frontal surface.

Weather and climate. Modern meteorology began when the daily weather map was developed as a device for analysis and forecasting, and the instrument that made this kind of map possible was the electromagnetic telegraph. In the United States the first telegraph line was strung in 1844 between Washington, D.C., and Baltimore. Concurrently with the expansion of telegraphic networks, the physicist Joseph Henry arranged for telegraph companies to have meteorological instruments in exchange for current data on weather telegraphed to the Smithsonian Institution. Some 500 stations had joined this cooperative effort by 1860. The Civil War temporarily prevented further expansion, but, meanwhile, a disaster of a different order had accelerated development of synoptic meteorology in Europe. On Nov. 14, 1854, an unexpected storm wrecked British and French warships off Balaklava on the Crimean peninsula. Had word of the approaching storm been telegraphed to this port in the Black Sea, the ships might have been saved. This mischance led in 1856 to the establishment of a national storm-warning service in France. In 1863 the Paris Observatory began publishing the first weather maps in modern format.

The first national weather service in the United States began operations in 1871 as an agency of the Department of War. The initial objective was to provide storm warnings for the Gulf and Atlantic coasts and the Great Lakes. In 1877 forecasts of temperature changes and precipitation averaged 74 percent in accuracy, as compared with 79 percent for cold-wave warnings. After 1878 daily weather maps were published.

Synoptic meteorology made possible the tracking of storm systems over wide areas. In 1868 the British meteorologist Alexander Buchan published a map showing the travels of a cyclonic depression across North America, the Atlantic, and into northern Europe. In the judgment of Sir Napier Shaw, Buchan's study marks the entry of modern meteorology, with "the weather map as its main feature and forecasting its avowed object."

Mapping regional variations

In addition to weather maps, a variety of other kinds of maps showing regional variations in the components of weather and climate were produced. In 1817 Alexander von Humboldt published a map showing the distribution of mean annual temperatures over the greater part of the Northern Hemisphere. Humboldt was the first to use isothermal lines in mapping temperature. Buchan drew the first maps of mean monthly and annual pressure for the entire world. Published in 1869, these maps added much to knowledge of the general circulation of the atmosphere. In 1886 Léon-Philippe Teisserenc de Bort of France published maps showing mean annual cloudiness over the Earth for each month and the year. The first world map of precipitation showing mean annual precipitation by isohyets was the work of Loomis in 1882. This work was further refined in 1899 by the maps of the British cartographer Andrew John Herbertson, which showed precipitation for each month of the year.

Although the 19th century was still in the age of meteorologic and climatological exploration, broad syntheses of old information thus kept pace with acquisition of the new fairly well. For example, Julius Hann's massive *Handbuch der Klimatologie* ("Handbook of Climatology"), first issued in 1883, is mainly a compendium of works published in the *Meteorologische Zeitschrift* ("Journal of Meteorology"). The *Handbuch* was kept current in revised editions until 1911, and this work is still sometimes called the most skillfully written account of world climate.

THE 20TH CENTURY: MODERN TRENDS AND DEVELOPMENTS

The impact of technological advances and the theory of plate tectonics

Geologic sciences. The development of the geologic sciences in the 20th century has been influenced by two major "revolutions." The first involves dramatic technological advances that have resulted in vastly improved instrumentation, the prime examples being the many types of highly sophisticated computerized devices. The second is centred on the development of the plate tectonics theory, which is the most profound and influential conceptual advance the Earth sciences have ever known.

Modern technological developments have affected all the different geologic disciplines. Their impact has been particularly notable in such activities as radiometric dating, experimental petrology, crystallography, chemical analysis of rocks and minerals, micropaleontology, and seismological exploration of the Earth's deep interior.

Radiometric dating. In 1905, shortly after the discovery of radioactivity, the American chemist Bertram Boltwood suggested that lead is one of the disintegration products of uranium, in which case the older a uranium-bearing mineral the greater should be its proportional part of lead. Analyzing specimens whose relative geologic ages were known, Boltwood found that the ratio of lead to uranium did indeed increase with age. After estimating the rate of this radioactive change he calculated that the absolute ages of his specimens ranged from 410,000,000 to 2,200,000,000 years. Though his figures were too high by about 20 percent, their order of magnitude was enough to dispose of the short scale of geologic time proposed by Lord Kelvin.

Versions of the modern mass spectrometer were invented in the early 1920s and 1930s, and during World War II the device was improved substantially to help in the development of the atomic bomb. Soon after the war, Harold C. Urey and G.J. Wasserburg applied the mass spectrometer to the study of geochronology. This device separates the different isotopes of the same element and can measure the variations in these isotopic abundances to within one part in 10,000. By determining the amount of the parent and daughter isotopes present in a sample and by knowing their rate of radioactive decay (each radioisotope has its own decay constant), the isotopic age of the sample can be calculated. For dating minerals and rocks, investigators commonly use the following couplets of parent and daughter isotopes: thorium-232–lead-208, uranium-235–lead-207, samarium-147–neodymium-143, rubidium-87–strontium-87, potassium-40–argon-40, and argon-40–argon-39.

Application of the mass spectrometer to geochronology

Such techniques have had an enormous impact on scientific knowledge of Earth history because precise dates can now be obtained on rocks in all orogenic (mountain) belts ranging in age from the early Archean (about 3,800,000,000 years old) to the late Tertiary (roughly 20,000,000 years old). The oldest sedimentary and igneous rocks in the world are found at Isua in western Greenland: they have an isotopic age of approximately 3,800,000,000 years—a fact first established in 1972 by Stephen Moorbath of the University of Oxford. Also by extrapolating backward in time to a situation when there was no lead that had been produced by radiogenic processes, a figure of about 4,600,000,000 years is obtained for the minimum age of the Earth. This figure is of the same order as ages obtained for certain meteorites and lunar rocks.

Experimental study of rocks. Experimental petrology began with the work of Jacobus Henricus van't Hoff, one of the founders of physical chemistry. Between 1896 and 1908 he elucidated the complex sequence of chemical reactions attending the precipitation of salts (evaporites) from the evaporation of seawater. Van't Hoff's aim was to explain the succession of mineral salts present in Permian rocks of Germany. His success at producing from aqueous solutions artificial minerals and rocks like those found in natural salt deposits stimulated studies of minerals crystallizing from silicate melts simulating the magmas from which igneous rocks have formed. Working at the Geo-

physical Laboratory of the Carnegie Institution of Washington, D.C., Norman L. Bowen conducted extensive phase-equilibrium studies of silicate systems, brought together in his *Evolution of the Igneous Rocks* (1928). Experimental petrology, both at the low-temperature range explored by van't Hoff and in the high ranges of temperature investigated by Bowen, continues to provide laboratory evidence for interpreting the chemical history of sedimentary and igneous rocks. Experimental petrology also provides valuable data on the stability limits of individual metamorphic minerals and of the reactions between different minerals in a wide variety of chemical systems. These experiments are carried out at elevated temperatures and pressures that simulate those operating in different levels of the Earth's crust. Thus the metamorphic petrologist today can compare the minerals and mineral assemblages found in natural rocks with comparable examples produced in the laboratory, the pressure–temperature limits of which have been well defined by experimental petrology.

Another branch of experimental science relates to the deformation of rocks. In 1906 the American physicist P.W. Bridgman developed a technique for subjecting rock samples to high pressures similar to those deep in the Earth. Studies of the behaviour of rocks in the laboratory have shown that their strength increases with confining pressure but decreases with rise in temperature. Down to depths of a few kilometres the strength of rocks would be expected to increase. At greater depths the temperature effect should become dominant, and response to stress should result in flow rather than fracture of rocks. In 1959 two American geologists, Marion King Hubbert and William W. Rubey, demonstrated that fluids in the pores of rock may reduce internal friction and permit gliding over nearly horizontal planes of the large overthrust blocks associated with folded mountains. More recently the Norwegian petrologist Hans Ramberg performed many experiments with a large centrifuge that produced a negative gravity effect and thus was able to create structures simulating salt domes, which rise because of the relatively low density of the salt in comparison with that of surrounding rocks. With all these deformation experiments, it is necessary to scale down as precisely as possible variables such as the time and velocity of the experiment and the viscosity and temperature of the material from the natural to the laboratory conditions.

Simulation of salt domes

Crystallography. In the 19th century crystallographers were only able to study the external form of minerals, and it was not until 1895 when the German physicist Wilhelm Conrad Röntgen discovered X rays that it became possible to consider their internal structure. In 1912 another German physicist, Max von Laue, realized that X rays were scattered and deflected at regular angles when they passed through a copper sulfate crystal, and so he produced the first X-ray diffraction pattern on a photographic film. A year later William Bragg of Britain and his son Lawrence perceived that such a pattern reflects the layers of atoms in the crystal structure, and they succeeded in determining for the first time the atomic crystal structure of the mineral halite (sodium chloride). These discoveries had a long-lasting influence on crystallography because they led to the development of the X-ray powder diffractometer, which is now widely used to identify minerals and to ascertain their crystal structure.

The chemical analysis of rocks and minerals. The successive stages in the plate-tectonic cycle (see below) produce igneous rocks that have geochemical patterns distinctive for each stage. Consequently, the igneous petrologist is able to use both the major and trace element abundances of rocks to define the possible tectonic environment in which they formed. The metamorphic petrologist can use the bulk composition of a recrystallized rock to define the composition of the original rock, assuming that there has been no change in composition during the metamorphic process. Advanced analytic geochemical equipment now make these studies possible.

Micropaleontology. Microscopic fossils, such as ostracods, foraminifera, and pollen grains, are common in sediments of the Mesozoic and Cenozoic eras (from about 225,000,000 years ago to the present). Because the rock chips brought up in oil wells are so small, a high-resolution instrument known as a scanning electron microscope had to be developed to study the microfossils. The classification of microfossils of organisms that lived within relatively short time spans has enabled Mesozoic-Cenozoic sediments to be subdivided in remarkable detail. This technique also has had a major impact on the study of Precambrian life (*i.e.*, organisms that existed more than 570,000,000 years ago). Carbonaceous spheroids and filaments about 7–10 millimetres (0.3–0.4 inch) long are recorded in 3,000,000,000-year-old sediments in the Pilbara region of northwestern Western Australia and in the lower Onverwacht Series of the Barberton belt in South Africa; these are the oldest reliable records of life on Earth.

Seismology and the structure of the Earth. Earthquake study was institutionalized in 1880 with the formation of the Seismological Society of Japan under the leadership of the English geologist John Milne. Milne and his associates invented the first accurate seismographs, including the instrument later known as the Milne seismograph. Seismology has revealed much about the structure of the Earth's core, mantle, and crust. The English seismologist Richard Dixon Oldham's studies of earthquake records in 1906 led to the discovery of the Earth's core. From studies of the Croatian quake of Oct. 8, 1909, the geophysicist Andrija Mohorovičić discovered the discontinuity (often called the Moho) that separates the crust from the underlying mantle.

Discovery of the Earth's core and of the Moho

Today there are more than 1,000 seismograph stations around the world, and their data are used to compile seismicity maps. These maps show that earthquake epicentres are aligned in narrow, continuous belts along the boundaries of lithospheric plates (see below). The earthquake foci outline the mid-oceanic ridges in the Atlantic, Pacific, and Indian oceans where the plates separate, while around the margins of the Pacific where the plates converge, they lie in a dipping plane, or Benioff zone, that defines the position of the subducting plate boundary to depths of about 700 kilometres.

Since 1950, additional information on the crust has been obtained from the analysis of artificial tremors produced by chemical explosions. These studies have shown that the Moho is present under all continents at an average depth of 35 kilometres and that the crust above it thickens under young mountain ranges to depths of 70 kilometres in the Andes and the Himalayas. In such investigations the reflections of the seismic waves generated from a series of "shot" points are also recorded, and this makes it possible to construct a profile of the subsurface structure. This is seismic reflection profiling, the main method of exploration used by the petroleum industry. During the late 1970s, a new technique for generating seismic waves was invented: thumping and vibrating the surface of the ground with a gas-propelled piston from a large truck.

The theory of plate tectonics. Plate tectonics has revolutionized virtually every discipline of the Earth sciences since the late 1960s and early 1970s. It has served as a unifying model or paradigm for explaining geologic phenomena that were formerly considered in unrelated fashion. Plate tectonics describes seismic activity, volcanism, mountain building, and various other Earth processes in terms of the structure and mechanical behaviour of a small number of enormous rigid plates thought to constitute the outer part of the planet (*i.e.*, the lithosphere). This all-encompassing theory grew out of observations and ideas about continental drift and seafloor spreading.

Plate tectonics as a unifying model

In 1912 the German meteorologist Alfred Wegener proposed that throughout most of geologic time there was only one continental mass, which he named Pangaea. At some time during the Mesozoic Era, Pangaea fragmented and the parts began to drift apart. Westward drift of the Americas opened the Atlantic Ocean, and the Indian block drifted across the Equator to join with Asia. In 1937 the South African Alexander Du Toit modified Wegener's hypothesis by suggesting the existence of two primordial continents: Laurasia in the north and Gondwanaland in the south. Aside from the congruency of continental shelf margins across the Atlantic, proponents of continental drift have amassed impressive geologic evidence to support their views. Similarities in fossil terrestrial organisms

The notion of continental drift

in pre-Cretaceous (older than 140,000,000 years) strata of Africa and South America and in pre-Jurassic rocks (older than 200,000,000 years) of Australia, India, Madagascar, and Africa are explained if these continents were formerly connected but difficult to account for otherwise. Fitting the Americas with the continents across the Atlantic brings together similar kinds of rocks and structures. Evidence of widespread glaciation during the Upper Paleozoic is found in Antarctica, southern South America, southern Africa, India, and Australia. If these continents were formerly united around the south polar region, this glaciation becomes explicable as a unified sequence of events in time and space.

Interest in continental drift heightened during the 1950s as knowledge of the Earth's magnetic field during the geologic past developed from the studies of Stanley K. Runcorn, Patrick M.S. Blackett, and others. Ferromagnetic minerals such as magnetite acquire a permanent magnetization when they crystallize as components of igneous rock. The direction of their magnetization is the same as the direction of the Earth's magnetic field at the place and time of crystallization. Particles of magnetized minerals released from their parent igneous rocks by weathering may later realign themselves with the existing magnetic field at the time these particles are incorporated into sedimentary deposits. Studies of the remanent magnetism in suitable rocks of different ages from over the world indicate that the magnetic poles were in different places at different times. The polar wandering curves are different for the several continents, but in important instances these differences are reconciled on the assumption that continents now separated were formerly joined. The curves for Europe and North America, for example, are reconciled by the assumption that America has drifted about 30° westward relative to Europe since the Triassic Period (195,000,000 to 230,000,000 years ago).

Remanent magnetism and polar wandering

In the early 1960s a major breakthrough in understanding the way the modern Earth works came from two studies of the ocean floor. First, the American geophysicists Harry H. Hess and Robert S. Dietz suggested that new ocean crust was formed along mid-oceanic ridges between separating continents; and second, Drummond H. Matthews and Frederick J. Vine of Britain proposed that the new oceanic crust acted like a magnetic tape recorder insofar as magnetic anomaly strips parallel to the ridge had been magnetized alternately in normal and reversed order, reflecting the changes in polarity of the Earth's magnetic field. This theory of seafloor spreading then needed testing, and the opportunity arose from major advances in deep-water drilling technology. The Joint Oceanographic Institutions Deep Earth Sampling (JOIDES) project began in 1969, continued with the Deep Sea Drilling Project (DSDP), and, since 1976, with the International Phase of Ocean Drilling (IPOD) project. These projects have produced more than 500 boreholes in the floor of the world's oceans, and the results have been as outstanding as the plate-tectonic theory itself. They confirm that the oceanic crust is everywhere younger than about 200,000,000 years and that the stratigraphic age determined by micropaleontology of the overlying oceanic sediments is close to the age of the oceanic crust calculated from the magnetic anomalies.

Theory of seafloor spreading

The plate-tectonic theory, which embraces both continental drift and seafloor spreading, was formulated in the mid-1960s by the Canadian geologist J. Tuzo Wilson, who described the network of mid-oceanic ridges, transform faults, and subduction zones as boundaries separating an evolving mosaic of enormous plates, and who proposed the idea of the opening and closing of oceans and eventual production of an orogenic belt by the collision of two continents.

Up to this point, no one had considered in any detail the implications of the plate-tectonic theory for the evolution of continental orogenic belts; most thought had been devoted to the oceans. In 1969 John Dewey of the University of Cambridge outlined an analysis of the Caledonian-Appalachian orogenic belts in terms of a complete plate-tectonic cycle of events, and this provided a model for the interpretation of other pre-Mesozoic (Paleozoic and Precambrian) belts. For a detailed discussion of plate-tectonic theory and its far-reaching effects, see PLATE TECTONICS.

Hydrologic sciences. *Water resources and seawater chemistry.* Quantitative studies of the distribution of water have revealed that an astonishingly small part of the Earth's water is contained in lakes and rivers. Ninety-seven percent of all the water is in the oceans; and, of the fresh water constituting the remainder, three-fourths is locked up in glacial ice and most of the rest is in the ground. Approximate figures are also now available for the amounts of water involved in the different stages of the hydrologic cycle. Of the 859 millimetres of annual global precipitation, 23 percent falls on the lands; but only about a third of the precipitation on the lands runs directly back to the sea, the remainder being recycled through the atmosphere by evaporation and transpiration. Subsurface groundwater accumulates by infiltration of rainwater into soil and bedrock. Some may run off into rivers and lakes, and some may reemerge as springs or aquifers. Advanced techniques are used extensively in groundwater studies nowadays. The rate of groundwater flow, for example, can be calculated from the breakdown of radioactive carbon-14 by measuring the time it takes for rainwater to pass through the ground, while numerical modeling is used to study heat and mass transfer in groundwater. High-precision equipment is used for measuring down-hole temperature, pressure, flow rate, and water level. Groundwater hydrology is important in studies of fractured reservoirs, subsidence resulting from fluid withdrawal, geothermal resource exploration, radioactive waste disposal, and aquifer thermal-energy storage.

Chemical analyses of trace elements and isotopes of seawater are conducted as part of the Geochemical Ocean Sections (Geosecs) program. Of the 92 naturally occurring elements, nearly 80 have been detected in seawater or in the organisms that inhabit it, and it is thought to be only a matter of time until traces of the others are detected. Contrary to the idea widely circulated in the older literature of oceanography, that the relative proportions of the oceans' dissolved constituents are constant, investigations since 1962 have revealed statistically significant variations in the ratios of calcium and strontium to chlorinity. The role of organisms as influences on the composition of seawater has become better understood with advances in marine biology. It is now known that plants and animals may collect certain elements to concentrations as much as 100,000 times their normal amounts in seawater. Abnormally high concentrations of beryllium, scandium, chromium, and iodine have been found in algae; of copper and arsenic in both the soft and skeletal parts of invertebrate animals; and of zirconium and cerium in plankton.

Desalinization, tidal power, and minerals from the sea. For ages a source of food and common salt, the sea is increasingly becoming a source of water, chemicals, and energy. In 1967 Key West, Fla., became the first U.S. city to be supplied solely by water from the sea, drawing its supplies from a plant that produces more than 2,000,000 gallons of refined water daily. Magnesia was extracted from the Mediterranean in the late 19th century; at present nearly all the magnesium metal used in the United States is mined from the sea at Freeport, Texas. Many ambitious schemes for using tidal power have been devised, but the first major hydrographic project of this kind was not completed until 1967, when a dam and electrical generating equipment were installed across the Rance River in Brittany. The seafloor and the strata below the continental shelves are also sources of mineral wealth. Concretions of manganese oxide, evidently formed in the process of subaqueous weathering of volcanic rocks, have been found in dense concentrations with a total abundance of 10^{11} tons. In addition to the manganese, these concretions contain copper, nickel, cobalt, zinc, and molybdenum. To date, oil and gas have been the most valuable products to be produced from beneath the sea.

Economic potential of the seafloor

Ocean bathymetry. Modern bathymetric charts show that about 20 percent of the surfaces of the continents are submerged to form continental shelves. Altogether the shelves form an area about the size of Africa. Continental slopes, which slant down from the outer edges of the

shelves to the abyssal plains of the seafloor, are nearly everywhere furrowed by submarine canyons. The depths to which these canyons have been cut below sea level seem to rule out the possibility that they are drowned valleys cut by ordinary streams. More likely the canyons were eroded by turbidity currents, dense mixtures of mud and water that originate as mud slides in the heads of the canyons and pour down their bottoms.

Profiling of the Pacific Basin prior to and during World War II resulted in the discovery of hundreds of isolated eminences rising 1,000 or more metres above the floor. Of particular interest were seamounts in the shape of truncated cones, whose flat tops rise to between 1.6 kilometres and a few hundred metres below the surface. Harry H. Hess interpreted the flat-topped seamounts (guyots) as volcanic mountains planed off by action of waves before they subsided to their present depths. Subsequent drilling in guyots west of Hawaii confirmed this view; samples of rocks from the tops contained fossils of Cretaceous age representing reef-building organisms of the kind that inhabit shallow water.

Ocean circulation, currents, and waves. Early in the century Vilhelm Bjerknes, a Norwegian meteorologist, and V. Walfrid Ekman, a Swedish physical oceanographer, investigated the dynamics of ocean circulation and developed theoretical principles that have influenced subsequent studies of currents in the sea. Bjerknes showed that very small forces resulting from pressure differences caused by nonuniform density of seawater can initiate and maintain fluid motion. Ekman analyzed the influence of winds and the Earth's rotation on currents. He theorized that in a homogeneous medium the frictional effects of winds blowing across the surface would cause movement of successively lower layers of water, the deeper the currents so produced the less their velocity and the greater their deflection by the Coriolis effect (an apparent force due to the Earth's rotation that causes deflection of a moving body to the right in the Northern Hemisphere and to the left in the Southern Hemisphere), until at some critical depth an induced current would move in a direction opposite to that of the wind.

Results of many investigations suggest that the forces that drive the ocean currents originate at the interface between water and air. The direct transfer of momentum from the atmosphere to the sea is doubtless the most important driving force for currents in the upper parts of the ocean. Next in importance are differential heating, evaporation, and precipitation across the air–sea boundary, altering the density of seawater and thus initiating movement of water masses with different densities. Studies of the properties and motion of water at depth have shown that strong currents also exist in the deep sea and that distinct types of water travel far from their geographic sources. For example, the highly saline water of the Mediterranean that flows through the Strait of Gibraltar has been traced over a large part of the Atlantic, where it forms a deep-water stratum that is circulated far beyond that ocean in currents around Antarctica.

Countercurrents and the Gulf Stream

Improvements in devices for determining the motion of seawater in three dimensions have led to the discovery of new currents and to the disclosure of unexpected complexities in the circulation of the oceans generally. In 1951 a huge countercurrent moving eastward across the Pacific was found below depths as shallow as 20 metres, and in the following year an analogous equatorial undercurrent was discovered in the Atlantic. In 1957 a deep countercurrent was detected beneath the Gulf Stream with the aid of subsurface floats emitting acoustic signals.

Since the 1970s Earth-orbiting satellites have yielded much information on the temperature distribution and thermal energy of ocean currents such as the Gulf Stream. Chemical analyses from Geosecs makes possible the determination of circulation paths, speeds, and mixing rates of ocean currents.

Surface waves of the ocean are also exceedingly complex, at most places and times reflecting the coexistence and interferences of several independent wave systems. During World War II, interest in forecasting wave characteristics was stimulated by the need for this critical information in the planning of amphibious operations. The oceanographers H.U. Sverdrup and Walter Heinrich Munk combined theory and empirical relationships in developing a method of forecasting "significant wave height"—the average height of the highest third of the waves in a wave train. Subsequently, this method was improved to permit wave forecasters to predict optimal routes for mariners. Forecasting of the most destructive of all waves, tsunamis, or "tidal waves," caused by submarine quakes and volcanic eruptions, is another recent development. Soon after 159 persons were killed in Hawaii by the tsunami of 1946, the U.S. Coast and Geodetic Survey established a seismic seawave warning system. Using a seismic network to locate epicentres of submarine quakes, the installation predicts the arrival of tsunamis at points around the Pacific Basin often hours before the arrival of the waves.

Glacier motion and the high-latitude ice sheets. Beginning around 1948, principles and techniques in metallurgy and solid-state physics were brought to bear on the mechanics of glacial movements. Laboratory studies showed that glacial ice deforms like other crystalline solids (such as metals) at temperatures near the melting point. Continued stress produces permanent deformation. In addition to plastic deformation within a moving glacier, the glacier itself may slide over its bed by mechanisms involving pressure melting and refreezing and accelerated plastic flow around obstacles. The causes underlying changes in rate of glacial movement, in particular spectacular accelerations called surges, require further study. Surges involve massive transfer of ice from the upper to the lower parts of glaciers at rates of as much as 20 metres a day, in comparison with normal advances of a few metres a year.

As a result of numerous scientific expeditions into Greenland and Antarctica, the dimensions of the remaining great ice sheets are fairly well known from gravimetric and seismic surveys. In parts of both continents it has been determined that the base of the ice is below sea level, probably due at least in part to subsidence of the crust under the weight of the caps. In 1966 a borehole was drilled 1,390 metres to bedrock on the North Greenland ice sheet, and two years later a similar boring of 2,162 metres was cut through the Antarctic ice at Byrd Station. From the study of annual incremental layers and analyses of oxygen isotopes, the bottom layers of ice cored in Greenland were estimated to be more than 150,000 years old, compared with 100,000 years for the Antarctic core. With the advent of geochemical dating of rocks it has become evident that the Ice Age, which in the earlier part of the century was considered to have transpired during the Quaternary Period, actually began much earlier. In Antarctica, for example, potassium–argon age determinations of lava overlying glaciated surfaces and sedimentary deposits of glacial origin show that glaciers existed on this continent at least 10,000,000 years ago.

Analysis of cores from the Greenland and Antarctic ice sheets

The study of ice sheets has benefited much from data produced by advanced instruments, computers, and orbiting satellites. The shape of ice sheets can be determined by numerical modeling, their heat budget from thermodynamic calculations, and their thickness with radar techniques. Colour images from satellites show the temperature distribution across the polar regions, which can be compared with the distribution of land and sea ice.

Atmospheric sciences. *Probes, satellites, and data transmission.* Kites equipped with meteorgraphs were used as atmospheric probes in the late 1890s, and in 1907 the U.S. Weather Bureau recorded the ascent of a kite to 7,044 metres above Mt. Weather, Virginia.

In the 1920s the radio replaced the telegraph and telephone as the principal instrument for transmitting weather data. By 1936 the radio meteorgraph (radiosonde) was developed, with capabilities of sending signals on relative humidity, temperature, and barometric pressure from unmanned balloons. Experimentation with balloons up to altitudes of about 31 kilometres showed that columns of warm air may rise more than 1.6 kilometres above the Earth's surface and that the lower atmosphere is often stratified, with winds in the different layers blowing in different directions. During the 1930s airplanes began to be used for observations of the weather, and the years since

1945 have seen the development of rockets and weather satellites. TIROS (Television Infra-Red Observation Satellite), the world's first all-weather satellite, was launched in 1960, and in 1964 the Nimbus Satellite of the United States National Aeronautics and Space Administration (NASA) was rocketed into near-polar orbit.

There are two types of weather satellites: polar and geostationary. Polar satellites, like Nimbus, orbit the Earth at low altitudes of a few hundred kilometres, and, because of their progressive drift, they produce a photographic coverage of the entire Earth every 24 hours. Geostationary satellites, first sent up in 1966, are situated over the Equator at altitudes of about 35,000 kilometres and transmit data at regular intervals. Much information can be derived from the data collected by satellites. For example, wind speed and direction are measured from cloud trajectories, while temperature and moisture profiles of the atmosphere are calculated from infrared data.

Weather forecasting. Efforts at incorporating numerical data on weather into mathematical formulas that could then be used for forecasting were initiated early in the century at the Norwegian Geophysical Institute. Vilhelm Bjerknes and his associates at Bergen succeeded in devising equations relating the measurable components of weather, but their complexity precluded the rapid solutions needed for forecasting. Out of their efforts, however, came the polar front theory for the origin of cyclones and the now-familiar names of cold front, warm front, and stationary front for the leading edges of air masses (see below *Study of the horizontal structure of the atmosphere*).

Bjerknes' polar front theory of cyclones

In 1922 the British mathematician Lewis Fry Richardson demonstrated that the complex equations of the Norwegian school could be reduced to long series of simple arithmetic operations. With no more than the desk calculators and slide rules then available, however, the solution of a problem in procedure only raised a new one in manpower. In 1946 the mathematician John von Neumann and his fellow workers at the Institute for Advanced Study, in Princeton, N.J., began work on an electronic device to do the computation faster than the weather developed. Four years later the von Neumann group could claim that, given adequate data, their computer could forecast the weather as well as a weatherman. Present-day numerical weather forecasting is achieved with the help of advanced computer analysis (see below *Weather forecasting*).

Cloud physics. Studies of cloud physics have shown that the nuclei around which water condenses vary widely in their degree of concentration and areal distribution, ranging from six per cubic centimetre over the oceans to more than 4,000,000 per cubic centimetre in the polluted air of some cities. The droplets that condense on these foreign particles may be as small as 0.001 centimetre in diameter. Raindrops apparently may form directly from the coalescence of these droplets, as in the case of tropical rains, or in the temperate zones through the intermediary of ice crystals. According to the theory of Tor Bergson and Walter Findeisen, vapour freezing on ice crystals in the clouds enlarges the crystals until they fall. What finally hits the ground depends on the temperature of air below the cloud—if below freezing, snow; if above, rain.

The Van Allen radiation belts

Properties and structure of the atmosphere. Less than a year after the space age began with the launching of the Soviet Sputnik I in 1957, the U.S. satellite Explorer I was sent into orbit with a Geiger counter for measuring the intensity of cosmic radiation at different levels above the ground. At altitudes around 1,000 kilometres this instrument ceased to function due to saturation by charged particles. This and subsequent investigations showed that a zone of radiation encircles the world between about latitude 75° N and 75° S, with maximum intensities at 5,000 and 16,000 kilometres. Named after the American physicist James Van Allen, a leading investigator of this portion of the Earth's magnetosphere, these zones are responsive to events taking place on the Sun. The solar wind, a stream of atomic particles emanating from the Sun in all directions, seems to be responsible for the electrons entrapped in the Van Allen region as well as for the teardrop shape of the magnetosphere as a whole, with its tail pointing always away from the Sun.

In 1898 Teisserenc de Bort, studying variations of temperature at high altitudes with the aid of balloons, discovered that at elevations of about 11 kilometres the figure for average decrease of temperature with height (about 5.5° C per 1,000 metres of ascent) dropped and the value remained nearly constant at around −55° C. He named the atmospheric zones below and above this temperature boundary the troposphere and the stratosphere.

Toward the end of World War II the B-29 Superfortress came into use as the first large aircraft to cruise regularly at 10,000 metres. Heading westward from bases in the Pacific, these planes sometimes encountered unexpected head winds that slowed their flight by as much as 300 kilometres per hour. The jet streams, as these high-altitude winds were named, have been found to encircle the Earth following wavy courses and moving from west to east at velocities ranging upward to 500 kilometres per hour. Aircraft have also proved useful in studies of the structure and dynamics of tropical hurricanes. Following the destruction wrought to the Atlantic Coast of the United States in 1955 by hurricanes Connie and Diane, a national centre was established in Florida with the missions of locating and tracking and, it is hoped, of learning how to predict the paths of hurricanes and to dissipate their energy.

Discovery of the jet streams

Weather modification. As late as the 1890s experiments were conducted in the United States in the hope of producing rain by setting off charges of dynamite lofted by balloons or kites. No positive results were reported, however. More promising were the cloud-seeding experiments of the 1940s, in which silver iodide was released into clouds as smoke or solid carbon dioxide broadcast into clouds from airplanes. The results are still uncertain for increasing precipitation. The lessons learned from cloud seeding, however, have had other successful applications, such as the dispersal of low-level supercooled fog at airports (the first system designed for this purpose, the Turboclair fog-dissipation system, was set up in 1970 at Orly airport in Paris).

The inadvertent weather modification that has followed industrialization and the building of large cities has, however, already produced measurable changes in local climate and may someday produce effects more widespread. The introduction of some 12,000,000,000 tons of carbon dioxide into the atmosphere each year from the burning of fuels may in time raise the Earth's average temperature. Cities affect the flow of wind, warm the atmosphere over them, and send pollutants into the sky. Updrafts and an abundance of condensation nuclei may increase rainfall and winter fog and reduce sunshine and daylight.

(C.C.A./B.F.W.)

Geologic sciences

An introduction to the geochemical and geophysical sciences logically begins with mineralogy because the Earth's rocks are composed of minerals—inorganic elements or compounds that have a fixed chemical composition and that are made up of regularly aligned rows of atoms. Today, one of the principal concerns of mineralogy is the chemical analysis of the some 3,000 known minerals that are the chief constituents of the three different rock types: sedimentary (formed by diagenesis of sediments deposited by surface processes); igneous (crystallized from magmas either at depth or at the surface as lavas); and metamorphic (formed by a recrystallization process at temperatures and pressures in the Earth's crust high enough to destabilize the parent sedimentary or igneous material). Geochemistry is the study of the composition of these different types of rocks.

Scope of the geologic sciences and the principal component disciplines

During mountain building, rocks became highly deformed, and the primary objective of structural geology is to elucidate the mechanism of formation of the many types of structures (*e.g.*, folds and faults) that arise from such deformation. The allied field of geophysics has several subdisciplines, which make use of different instrumental techniques. Seismology, for example, involves the exploration of the Earth's deep structure through the detailed analysis of recordings of elastic waves generated by earthquakes and man-made explosions. Earthquake seismology

has largely been responsible for defining the location of major plate boundaries and of the dip of subduction zones down to depths of about 700 kilometres at those boundaries. In other subdisciplines of geophysics, gravimetric techniques are used to determine the shape and size of underground structures; electrical methods help to locate a variety of mineral deposits that tend to be good conductors of electricity; and paleomagnetism has played the principal role in tracking the drift of continents.

Geomorphology

Geomorphology is concerned with the surface processes that create the landscapes of the world—namely, weathering and erosion. Weathering is the alteration and breakdown of rocks at the Earth's surface caused by local atmospheric conditions, while erosion is the process by which the weathering products are removed by water, ice, and wind. The combination of weathering and erosion leads to the wearing down or denudation of mountains and continents, with the erosion products being deposited in rivers, internal drainage basins, and the oceans. Erosion is thus the complement of deposition. The unconsolidated accumulated sediments are transformed by the process of diagenesis and lithification into sedimentary rocks, thereby completing a full cycle of the transfer of matter from an old continent to a young ocean and ultimately to the formation of new sedimentary rocks. Knowledge of the processes of interaction of the atmosphere and the hydrosphere with the surface rocks and soils of the Earth's crust is important for an understanding not only of the development of landscapes but also (and perhaps more importantly) of the ways in which sediments are created. This in turn helps in interpreting the mode of formation and the depositional environment of sedimentary rocks. Thus the discipline of geomorphology is fundamental to the uniformitarian approach to the Earth sciences according to which the present is the key to the past.

Geologic history provides a conceptual framework and overview of the evolution of the Earth. An early development of the subject was stratigraphy, the study of order and sequence in bedded sedimentary rocks. Stratigraphers still use the two main principles established by the late 18th-century English engineer and surveyor William Smith, regarded as the father of stratigraphy: (1) that younger beds rest upon older ones and (2) different sedimentary beds contain different and distinctive fossils, enabling beds with similar fossils to be correlated over large distances. Today, biostratigraphy uses fossils to characterize successive intervals of geologic time, but as relatively precise time markers only to the beginning of the Cambrian Period, about 540,000,000 years ago. The geologic time scale, back to the oldest rocks, some 3,900,000,000 years ago, can be quantified by isotopic dating techniques. This is the science of geochronology, which in recent years has revolutionized scientific perception of Earth history and which relies heavily on the measured parent-to-daughter ratio of radiogenic isotopes (see below).

Paleontology is the study of fossils and is concerned not only with their description and classification but also with an analysis of the evolution of the organisms involved. Simple fossil forms can be found in early Precambrian rocks as old as 3,500,000,000 years, and it is widely considered that life on Earth must have begun before the appearance of the oldest rocks. Paleontological research of the fossil record since the Cambrian Period has contributed much to the theory of evolution of life on Earth.

Several disciplines of the geologic sciences have practical benefits for society. The geologist is responsible for the discovery of minerals, oil, gas, and coal, which are the main economic resources of the Earth; for the application of knowledge of subsurface structures and geologic conditions to the building industry; and for the prevention of natural hazards or at least providing early warning of their occurrence. (For further examples, see below *Practical applications*.)

Astrogeology is important in that it contributes to understanding the development of the Earth within the solar system. The U.S. Apollo program of manned missions to the Moon, for example, provided scientists with firsthand information on lunar geology, including observations on such features as meteorite craters that are relatively rare on Earth. Unmanned space probes have yielded significant data on the surface features of many of the planets and their satellites. Since the 1970s even such distant planetary systems as those of Jupiter, Saturn, and Uranus have been explored by probes.

STUDY OF THE COMPOSITION OF THE EARTH

Mineralogy. As a discipline, mineralogy has had close historical ties with geology. Minerals as basic constituents of rocks and ore deposits are obviously an integral aspect of geology. The problems and techniques of mineralogy, however, are distinct in many respects from those of the rest of geology, with the result that mineralogy has grown to be a large, complex discipline in itself.

About 3,000 distinct mineral species are recognized, but relatively few are important in the kinds of rocks that are abundant in the outer part of the Earth. Thus a few minerals such as the feldspars, quartz, and mica are the essential ingredients in granite and its near relatives. Limestones, which are widely distributed on all continents, consist largely of only two minerals, calcite and dolomite. Many rocks have a more complex mineralogy, and in some the mineral particles are so minute that they can be identified only through specialized techniques.

Identifying minerals on the basis of physical properties

It is possible to identify an individual mineral in a specimen by examining and testing its physical properties. Determining the hardness of a mineral is the most practical way of identifying it. This can be done by using the Mohs scale of hardness, which lists 10 common minerals in their relative order of hardness: talc (softest with the scale number 1), gypsum (2), calcite (3), fluorite (4), apatite (5), orthoclase (6), quartz (7), topaz (8), corundum (9), and diamond (10). Harder minerals scratch softer ones, so that an unknown mineral can be readily positioned between minerals on the scale. Certain common objects that have been assigned hardness values roughly corresponding to those of the Mohs scale (*e.g.*, fingernail [2.5], pocketknife blade [5.5], steel file [6.5]) are usually used in conjunction with the minerals on the scale for additional reference.

Other physical properties of minerals that aid in identification are crystal form, cleavage type, fracture, streak, lustre, colour, specific gravity, and density. In addition, the refractive index of a mineral can be determined with precisely calibrated immersion oils. Some minerals have distinctive properties that help to identify them. For example, carbonate minerals effervesce with dilute acids; halite is soluble in water and has a salty taste; fluorite (and about 100 other minerals) fluoresces in ultraviolet light; and uranium-bearing minerals are radioactive.

The science of crystallography is concerned with the geometric properties and internal structure of crystals. Because minerals are generally crystalline, crystallography is an essential aspect of mineralogy. Investigators in the field may use a reflecting goniometer that measures angles between crystal faces to help determine the crystal system to which a mineral belongs. Another instrument that they frequently employ is the X-ray diffractometer, which makes use of the fact that X rays, when passing through a mineral specimen, are diffracted at regular angles. The paths of the diffracted rays are recorded on photographic film, and the positions and intensities of the resulting diffraction lines on the film provide a particular pattern. Every mineral has its own unique diffraction pattern, so crystallographers are able to determine not only the crystal structure of a mineral but the type of mineral as well.

When a complex substance such as a magma crystallizes to form igneous rock, the grains of different constituent minerals grow together and mutually interfere, with the result that they do not retain their externally recognizable crystal form. To study the minerals in such a rock, the mineralogist uses a petrographic microscope constructed for viewing thin sections of the rock, which are ground uniformly to a thickness of about 0.03 millimetre, in light polarized by two polarizing prisms in the microscope. If the rock is crystalline, its essential minerals can be determined by their peculiar optical properties as revealed in transmitted light under magnification, provided that the individual crystal grains can be distinguished. Opaque minerals, such as those with a high content of metallic

elements, require a technique employing reflected light from polished surfaces. This kind of microscopic analysis has particular application to metallic ore minerals. The polarizing microscope, however, has a lower limit to the size of grains that can be distinguished with the eye; even the best microscopes cannot resolve grains less than about 0.5 micrometre (0.0005 millimetre) in diameter. For higher magnifications the mineralogist uses an electron microscope, which produces images with diameters enlarged tens of thousands of times.

Studying the chemical composition of minerals

The methods described above are based on a study of the physical properties of minerals. Another important area of mineralogy is concerned with the chemical composition of minerals. The primary instrument used is the electron microprobe. Here, a beam of electrons is focused on a thin section of rock that has been highly polished and coated with carbon. The electron beam can be narrowed to a diameter of about one micrometre and thus can be focused on a single grain of a mineral, which can be observed with an ordinary optical-microscope system. The electrons cause the atoms in the mineral under examination to emit diagnostic X rays, the intensity and concentration of which are measured by a computer. Besides spot analysis, this method allows a mineral to be traversed for possible chemical zoning. Moreover, the concentration and relative distribution of elements such as magnesium and iron across the boundary of two coexisting minerals like garnet and pyroxene can be used with thermodynamic data to calculate the temperature and pressure at which minerals of this type crystallize.

Determining the origin of minerals

Although the major concern of mineralogy is to describe and classify the geometrical, chemical, and physical properties of minerals, it is also concerned with their origin. Physical chemistry and thermodynamics are basic tools for understanding mineral origin. Some of the observational data of mineralogy consist of the behaviour of solutions in precipitating crystalline materials under controlled conditions in the laboratory. Certain minerals can be created synthetically under conditions in which temperature and concentration of solutions are carefully monitored. Other experimental methods include study of the transformation of solids at high temperatures and pressures to yield specific minerals or assemblages of minerals. Experimental data obtained in the laboratory, coupled with chemical and physical theory, enable the conditions of origin of many naturally occurring minerals to be inferred.

Petrology. Petrology is the study of rocks, and, because most rocks are composed of minerals, petrology is strongly dependent on mineralogy. In many respects mineralogy and petrology share the same problems; for example, the physical conditions that prevail (pressure, temperature, time, and presence or absence of water) when particular minerals or mineral assemblages are formed. Although petrology is in principle concerned with rocks throughout the crust, as well as with those of the inner depths of the Earth, in practice the discipline deals mainly with those that are accessible in the outer part of the Earth's crust. Rock specimens obtained from the surface of the Moon and from other planets are also proper considerations of petrology. Fields of specialization in petrology correspond to the aforementioned three major rock types—igneous, sedimentary, and metamorphic.

Igneous petrology. Igneous petrology is concerned with the identification, classification, origin, evolution, and processes of formation and crystallization of the igneous rocks. Most of the rocks available for study come from the Earth's crust, but a few, such as eclogites, derive from the mantle. The scope of igneous petrology is very large because igneous rocks make up the bulk of the continental and oceanic crusts and of the mountain belts of the world, which range in age from early Archean to the late Tertiary Period; and they also include the high-level volcanic extrusive rocks and the plutonic rocks that formed deep within the crust. Of utmost importance to igneous petrologic research is geochemistry, which is concerned with the major- and trace-element composition of igneous rocks as well as of the magmas from which they arose. Some of the major problems within the scope of igneous petrology are: (1) the form and structure of igneous bodies, whether they be lava flows or granitic intrusions, and their relations to surrounding rocks (these are problems studied in the field); (2) the crystallization history of the minerals that make up igneous rocks (this is determined with the petrographic polarizing microscope); (3) the classification of rocks based on textural features, grain size, and the abundance and composition of constituent minerals; (4) the fractionation of parent magmas by the process of magmatic differentiation, which may give rise to an evolutionary sequence of genetically related igneous products; (5) the mechanism of generation of magmas by partial melting of the lower continental crust, suboceanic and subcontinental mantle, and subducting slabs of oceanic lithosphere; (6) the history of formation and the composition of the present oceanic crust determined on the basis of data from the International Phase of Ocean Drilling (IPOD) project; (7) the evolution of igneous rocks through geologic time; (8) the composition of the mantle from studies of the rocks and mineral chemistry of eclogites brought to the surface in kimberlite pipes; (9) the conditions of pressure and temperature at which different magmas form and at which their igneous products crystallize (determined from high-pressure experimental petrology).

Major concerns of igneous petrology

Instrumentation

The basic instrument of igneous petrology is the petrographic polarizing microscope, but the majority of instruments used today have to do with determining rock and mineral chemistry. These include the X-ray fluorescence spectrometer, equipment for neutron activation analysis, induction-coupled plasma spectrometer, electron microprobe, ionprobe, and mass spectrometer. These instruments are highly computerized and automatic and produce analyses rapidly (see below *Geochemistry*). Complex high-pressure experimental laboratories also provide vital data.

With a vast array of sophisticated instruments available, the igneous petrologist is able to answer many fundamental questions. Study of the ocean floor has been combined with investigation of ophiolite complexes, which are interpreted as slabs of ocean floor that have been thrust onto adjacent continental margins. An ophiolite provides a much deeper section through the ocean floor than is available from shallow drill cores and dredge samples from the extant ocean floor. These studies have shown that the topmost volcanic layer consists of tholeiitic basalt or mid-ocean ridge basalt that crystallized at an accreting rift or ridge in the middle of an ocean. A combination of mineral chemistry of the basalt minerals and experimental petrology of such phases allows investigators to calculate the depth and temperature of the magma chambers along the mid-ocean ridge. The depths are close to six kilometres, and the temperatures range from 1,150° C to 1,279° C. Comprehensive petrologic investigation of all the layers in an ophiolite makes it possible to determine the structure and evolution of the associated magma chamber.

In 1974 B.W. Chappell and A.J.R. White discovered two major and distinct types of granitic rock—namely, I- and S-type granitoids. The I-type has strontium-87/strontium-86 ratios lower than 0.706 and contains magnetite, sphene, and allanite but no muscovite. These rocks formed above subduction zones in island arcs and active (subducting) continental margins and were ultimately derived by partial melting of mantle and subducted oceanic lithosphere. In contrast, S-type granitoids have strontium-87/strontium-86 ratios higher than 0.706 and contain muscovite, ilmenite, and monazite. These rocks were formed by partial melting of lower continental crust. Those found in the Himalayas were formed during the Miocene Epoch some 20,000,000 years ago as a result of the penetration of India into Asia, which thickened the continental crust and then caused its partial melting.

In the island arcs and active continental margins that rim the Pacific Ocean, there are many different volcanic and plutonic rocks belonging to the calc-alkaline series. These include basalts, andesites, dacites, rhyolites, ignimbrites, tonalites, granodiorites, diorites, granites, peridotites, and gabbros. They occur typically in vast batholiths, which may reach several thousand kilometres in length and contain more than 1,000 separate granitic bodies. These calc-alkaline rocks represent the principal means of growth of the continental crust probably throughout the whole

of geologic time. Much research is devoted to them in an effort to determine the source regions of their parent magmas, the chemical evolution of the magmas, and the interrelationships of the resultant igneous rocks. Although there are still many disagreements between alternative models and different petrologists, it is generally agreed that the magmas were derived from the mantle and that there is a small chemical contribution from both the subducted oceanic slab and the continental crust through which the magmas rose. One of the major influences on the evolution of these rocks is the presence of water, which was derived originally from the dehydration of the subducted slab.

Sedimentary petrology. The field of sedimentary petrology is concerned with the description and classification of sedimentary rocks, interpretation of the processes of transportation and deposition of the sedimentary materials forming the rocks, the environment that prevailed at the time the sediments were deposited, and the alteration (compaction, cementation, and chemical and mineralogical modification) of the sediments after deposition.

Carbonate petrology and clastic petrology

There are two main branches of sedimentary petrology. One branch deals with carbonate rocks, namely limestones and dolomites, composed principally of calcium carbonate (calcite) and calcium magnesium carbonate (dolomite). Much of the complexity in classifying carbonate rocks stems partly from the fact that many limestones and dolomites have been formed, directly or indirectly, through the influence of organisms, including bacteria, lime-secreting algae, various shelled organisms (*e.g.,* mollusks and brachiopods), and by corals. In limestones and dolomites that were deposited under marine conditions, commonly in shallow warm seas, much of the material initially forming the rock consists of skeletons of lime-secreting organisms. In many examples, this skeletal material is preserved as fossils. Some of the major problems of carbonate petrology concern the physical and biological conditions of the environments in which carbonate material has been deposited, including water depth, temperature, degree of illumination by sunlight, motion by waves and currents, and the salinity and other chemical aspects of the water in which deposition occurred.

The other principal branch of sedimentary petrology is concerned with the sediments and sedimentary rocks that are essentially noncalcareous. These include sands and sandstones, clays and claystones, siltstones, conglomerates, glacial till, and varieties of sandstones, siltstones, and conglomerates (*e.g.,* the graywacke-type sandstones and siltstones). These rocks are broadly known as clastic rocks because they consist of distinct particles or clasts. Clastic petrology is concerned with classification, particularly with respect to the mineral composition of fragments or particles, as well as the shapes of particles (angular versus rounded), and the degree of homogeneity of particle sizes. Other main concerns of clastic petrology are the mode of transportation of sedimentary materials, including the transportation of clay, silt, and fine sand by wind; and the transportation of these and coarser materials through suspension in water, through traction by waves and currents in rivers, lakes, and seas, and sediment transport by ice.

Sedimentary petrology also is concerned with the small-scale structural features of sediments and sedimentary rocks. Features that can be conveniently seen in a specimen held in the hand are within the domain of sedimentary petrology. These features include the geometrical attitude of mineral grains with respect to each other, small-scale cross stratification, the shapes and interconnections of pore spaces, and the presence of fractures and veinlets.

Instruments and methods used by sedimentary petrologists include the petrographic microscope for description and classification, X-ray mineralogy for defining fabrics and small-scale structures, physical model flume experiments for studying the effects of flow as an agent of transport and the development of sedimentary structures, and mass spectrometry for calculating stable isotopes and the temperatures of deposition, cementation, and diagenesis. Wet-suit diving permits direct observation of current processes on coral reefs, and manned submersibles enable observation at depth on the ocean floor and in mid-oceanic ridges.

The plate-tectonic theory has given rise to much interest in the relationships between sedimentation and tectonics, particularly in modern plate-tectonic environments—*e.g.,* spreading-related settings (intracontinental rifts, early stages of intercontinental rifting such as the Red Sea, and late stages of intercontinental rifting such as the margins of the present Atlantic Ocean), mid-oceanic settings (ridges and transform faults), subduction-related settings (volcanic arcs, fore-arcs, back-arcs, and trenches), and continental collision-related settings (the Alpine-Himalayan belt and late orogenic basins with molasse [*i.e.,* thick association of clastic sedimentary rocks consisting chiefly of sandstones and shales]). Today, many subdisciplines of sedimentary petrology are concerned with the detailed investigation of the various sedimentary processes that occur within these plate-tectonic environments.

Metamorphic petrology. Metamorphism means change in form. In geology the term is used to refer to a solid-state recrystallization of earlier igneous, sedimentary, or metamorphic rocks. There are two main types of metamorphism: (1) contact metamorphism, in which changes induced largely by increase in temperature are localized at the contacts of igneous intrusions; and (2) regional metamorphism, in which increased pressure and temperature have caused recrystallization over extensive regions in mountain belts. Other types of metamorphism include local effects caused by deformation in fault zones, burning oil shales, and thrusted ophiolite complexes; extensive recrystallization caused by high heat flow in mid-ocean ridges; and shock metamorphism induced by high-pressure impacts of meteorites in craters on the Earth and Moon.

Types of metamorphism

Metamorphic petrology is concerned with field relations and local tectonic environments; the description and classification of metamorphic rocks in terms of their texture and chemistry, which provides information on the nature of the premetamorphic material; the study of minerals and their chemistry (the mineral assemblages and their possible reactions), which yields data on the temperatures and pressures at which the rocks recrystallized; and the study of fabrics and the relations of mineral growth to deformation stages and major structures, which provides information about the tectonic conditions under which regional metamorphic rocks formed.

A supplement to metamorphism is metasomatism: the introduction and expulsion of fluids and elements through rocks during recrystallization. When new crust is formed and metamorphosed at a mid-oceanic ridge, seawater penetrates into the crust for a few kilometres and carries much sodium with it. During formation of a contact metamorphic aureole around a granitic intrusion, hydrothermal fluids carrying elements such as iron, boron, and fluorine pass from the granite into the wall rocks. When the continental crust is thickened, its lower part may suffer dehydration and form granulites. The expelled fluids, carrying such heat-producing elements as rubidium, uranium, and thorium migrate upward into the upper crust. Much petrologic research is concerned with determining the amount and composition of fluids that have passed through rocks during these metamorphic processes.

The basic instrument used by the metamorphic petrologist is the petrographic microscope, which allows detailed study and definition of mineral types, assemblages, and reactions. If a heating/freezing stage is attached to the microscope, the temperature of formation and composition of fluid inclusions within minerals can be calculated. These inclusions are remnants of the fluids that passed through the rocks during the final stages of their recrystallization. The electron microprobe is widely used for analyzing the composition of the component minerals. The petrologist can combine the mineral chemistry with data from experimental studies and thermodynamics to calculate the pressures and temperatures at which the rocks recrystallized. By obtaining information on the isotopic age of successive metamorphic events with a mass spectrometer, pressure–temperature–time curves can be worked out. These curves chart the movement of the rocks over time as they were brought to the surface from deep within the continental crust. This technique is important for understanding metamorphic processes in two plate-tectonic

environments: (1) in subduction zones where rocks are carried down to several tens of kilometres and are then brought up to be exposed at the present surface; and (2) in thrust zones in mountain belts produced by continental collisions (*e.g.*, the Himalayas), where deep crustal rocks have been brought up to high levels by major thrusts. These examples demonstrate that metamorphic petrology plays a key role in unraveling tectonic processes in mountain belts that have passed through the plate-tectonic cycle of events.

Economic geology. The mineral commodities on which modern civilization is heavily dependent are obtained from the Earth's crust and have a prominent place in the study and practice of economic geology. In turn, economic geology consists of several principal branches that include the study of ore deposits, petroleum geology, and the geology of nonmetallic deposits (excluding petroleum), such as coal, stone, salt, gypsum, clay and sand, and other commercially valuable materials.

Branches and objectives of economic geology

The practice of economic geology is distinguished by the fact that its objectives are to aid in the exploration for and extraction of mineral resources. The objectives are therefore economic. In petroleum geology, for example, a common goal is to guide oil-well-drilling programs so that the most profitable prospects are drilled, and those that are likely to be of marginal economic value, or barren, are avoided. A similar philosophy influences the other branches of economic geology. In this sense, economic geology can be considered as an aspect of business that is devoted to economic decision making. Many deposits of economic interest, particularly those of metallic ores, are of extreme scientific interest in themselves, however, and they have warranted intensive study that has been somewhat apart from economic considerations.

The practice of economic geology provides employment for a large number of geologists. On a worldwide basis, probably more than two-thirds of those persons employed in the geologic sciences are engaged in work that touches on the economic aspects of geology. These include geologists whose main interests lie in diverse fields of the geologic sciences. For example, the petroleum industry, which collectively is the largest employer of economic geologists, attracts individuals with specialties in stratigraphy, sedimentary petrology, structural geology, paleontology, and geophysics.

Geochemistry. *Chemistry of the Earth.* Geochemistry is broadly concerned with the application of chemistry to virtually all aspects of geology. Inasmuch as the Earth is composed of the chemical elements, all geologic materials and most geologic processes can be regarded from a chemical point of view. Some of the major problems that broadly belong to geochemistry are as follows: the origin and abundance of the elements in the solar system, galaxy, and universe (cosmochemistry); the abundance of elements in the major divisions of the Earth, including the core, mantle, crust, hydrosphere, and atmosphere; the behaviour of ions in the structure of crystals; the chemical reactions in cooling magmas and the origin and evolution of deeply buried intrusive igneous rocks; the chemistry of volcanic (extrusive) igneous rocks and of phenomena closely related to volcanic activity, including hot-spring activity, emanation of volcanic gases, and origin of ore deposits formed by hot waters derived during the late stages of cooling of igneous magmas; chemical reactions involved in weathering of rocks in which earlier formed minerals decay and new minerals are created; the transportation of weathering products in solution by natural waters in the ground and in streams, lakes, and the sea; chemical changes that accompany compaction and cementation of unconsolidated sediments to form sedimentary rocks; and the progressive chemical and mineralogical changes that take place as rocks undergo metamorphism.

Kinds of geochemical problems

One of the leading general concerns of geochemistry is the continual recycling of the materials of the Earth. This process takes place in several ways: (1) It is widely believed that oceanic and continental basalts crystallized from magmas that were ultimately derived by partial melting of the Earth's mantle. Much geochemical research is devoted to the quantification of this extraction of mantle material and its contribution to crustal growth throughout geologic time in the many stages of seafloor formation and mountain building. (2) When the basalts that formed at the mid-oceanic ridge are transported across the ocean by the process of seafloor spreading, they interact with seawater, and this involves the adding of sodium to the basaltic crust and the extraction of calcium from it. (3) Geophysical data confirm the idea that the oceanic lithosphere is being consumed along the Earth's major subduction zones below the continental lithosphere—*e.g.*, along the continental margin of the Andes Mountain Ranges. This may involve pelagic sediments from the ocean floor, oceanic basalts altered by seawater exchange, gabbros, ultramafic rocks, and segments of the underlying mantle. Many geochemists are studying what happens to this subducted material and how it contributes to the growth of island arcs and Andean-type mountain belts. (4) The behaviour of dissolved materials in natural waters, under the relatively low temperatures that prevail at or near the surface of the Earth, is an integral aspect of the crustal cycle. Weathering processes supply dissolved material, including silica, calcium carbonate, and other salts, to streams. These materials then enter the oceans, where some remain in solution (*e.g.*, sodium chloride), whereas others are progressively removed to form certain sedimentary rocks, including limestone and dolomite, and, where conditions are conducive for the formation of deposits by means of evaporation, gypsum (hydrous calcium sulfate), rock salt (halite), and potash deposits may occur.

Organic geochemistry and biogeochemistry

The behaviour of biological materials and their subsequent disposition are important aspects of geochemistry, generally termed organic geochemistry and biogeochemistry. Major problems of organic geochemistry include the question of the chemical environment on Earth in which life originated; the modification of the hydrosphere, and particularly the atmosphere, through the effects of life; and the incorporation of organic materials in rocks, including carbonaceous material in sedimentary rocks. The nature and chemical transformations of biological material present in deposits of coal, petroleum, and natural gas lie within the scope of organic geochemistry. Organic chemical reactions influence many geochemical processes, as, for example, rock weathering and production of soil, the solution, precipitation, and secretion of such dissolved materials as calcium carbonate, and the alteration of sediments to form sedimentary rocks. Biogeochemistry deals chiefly with the cyclic flows of individual elements and their compounds between living and nonliving systems.

Geochemistry has applications to other subdisciplines within geology, as well as to disciplines relatively far removed from it. At one extreme, geochemistry is linked with cosmology in a number of ways. These include the study of the chemical composition of meteorites, the relative abundance of elements in the Earth, Moon, and other planets, and the ages of meteorites and of rocks of the crust of the Earth and Moon as established by radiometric means. At the other extreme, the geochemistry of traces of metals in rocks and soils and, ultimately, in the food chain has important consequences for humans and for the vast body of lesser organisms on which they are dependent and with whom they coexist. Deficiencies in traces of copper and cobalt in forage plants, for example, lead to diseases in certain grazing animals and may locally influence human health. These deficiencies are in turn related to the concentrations of these elements in rocks and the manner in which they are chemically combined within soils and rocks.

Instruments and techniques for the chemical analysis of rocks

The chemical analysis of minerals is undertaken with the electron microprobe (see above). Instruments and techniques used for the chemical analysis of rocks are as follows: The X-ray fluorescent (XRF) spectrometer excites atoms with a primary X-ray beam and causes secondary (or fluorescent) X rays to be emitted. Each element produces a diagnostic X radiation, the intensity of which is measured. This intensity is proportional to the concentration of the element in the rock, and so the bulk composition can be calculated. The crushed powder of the rock is compressed into a disk or fused into a bead and loaded into the spectrometer, which analyzes it automatically under computer

control. Analysis of most elements having concentrations of more than five parts per million is possible.

Neutron-activation analysis is based on the fact that certain elements are activated or become radiogenic when they are bombarded with a flux of neutrons formed from the radioactive decay of uranium-235 in a nuclear reactor. With the addition of the neutrons, the stable isotopes produce new unstable radionuclides, which then decay, emitting particles with diagnostic energies that can be separated and measured individually. The technique is particularly suitable for the analysis of the rare earth elements, uranium, thorium, barium, and hafnium, with a precision to less than one part per million.

The induction-coupled plasma (ICP) spectrometer can analyze over 40 elements. Here, a solution of a rock is put into a plasma, and the concentration of the elements is determined from the light emitted. This method is rapid, and the ICP spectrometer is particularly suited to analyzing large numbers of soil and stream sediment samples, as well as mineralized rocks in mineral exploration.

Isotopic geochemistry. Isotopic geochemistry has several principal roles in geology. One is concerned with the enrichment or impoverishment of certain isotopic species that results from the influence of differences in mass of molecules containing different isotopes. Measurements of the proportions of various isotopic species can be used as a form of geologic thermometer. The ratio of oxygen-16 to oxygen-18 in calcium carbonate secreted by various marine organisms from calcium carbonate in solution in seawater is influenced by the temperature of the seawater. Precise measurement of the proportions of oxygen-16 with respect to oxygen-18 in calcareous shells of some fossil marine organisms provides a means of estimating the temperatures of the seas in which they lived. The varying ocean temperatures during and between the major advances of glaciers during the ice ages have been inferred by analyzing the isotopic composition of the skeletons of floating organisms recovered as fossils in sediment on the seafloor. Other uses of isotopic analyses that involve temperature-dependent rate processes include the progressive removal of crystals from cooling igneous magmas.

Radiometric dating

Another role of isotopic geochemistry that is of great importance in geology is radiometric age dating. The ability to quantify the geologic time scale—*i.e.,* to date the events of the geologic past in terms of numbers of years—is largely a result of coupling radiometric-dating techniques with older, classical methods of establishing relative geologic ages. As explained earlier, radiometric-dating methods are based on the general principle that a particular radioactive isotope (radioactive parent or source material) incorporated in geologic material decays at a uniform rate, producing a decay product, or daughter isotope. Some radiometric "clocks" are based on the ratio of the proportion of parent to daughter isotopes, others on the proportion of parent remaining, and still others on the proportion of daughter isotopes with respect to each other. For example, uranium-238 decays ultimately to lead-206, which is one of the four naturally occurring isotopic species of lead. Minerals that contain uranium-238 when initially formed may be dated by measuring the proportions of lead-206 and uranium-238; the older the specimen, the greater the proportion of lead-206 with respect to uranium-238. The decay of potassium-40 to form argon-40 (calcium-40 is produced in this decay process as well) is also a widely used radiometric-dating tool, though there are several other parent-daughter pairs that are used in radiometric dating, including another isotope of uranium (uranium-235), which decays ultimately to form lead-207, and thorium-232, which decays to lead-208.

Uranium-238 and uranium-235 decay very slowly, although uranium-235 decays more rapidly than uranium-238. The rate of decay may be expressed in several ways. One way is by the radioactive isotope's half-life—the interval of time in which half of any given initial amount will have decayed. The half-life of uranium-238 is about 4,510,000,000 years, whereas the half-life of uranium-235 is about 713,000,000 years. Other radioactive isotopes decay at greatly differing rates, with half-lives ranging from a fraction of a second to quadrillions of years.

It is useful to combine a variety of isotopic methods to determine the complete history of a crustal rock. A samarium-147–neodymium-143 date on a granitic gneiss, for example, may be interpreted as the time of mantle–crust differentiation or crustal accretion that produced the original magmatic granite. Also, a lead-207–lead-206 date on a zircon will indicate the crystallization age of the granite. In contrast, a rubidium-87–strontium-87 date of a whole rock sample may give the time at which the rock became a closed system for migration of the strontium during the period of metamorphism that converted the granite to a granitic gneiss. When potassium-40 breaks down to argon-40, the argon continues to diffuse until the rock has cooled to about 200° C; therefore, a potassium-40–argon-40 date may be interpreted as the time when the granite cooled through a blocking temperature that stopped all argon release. This may reflect the cooling of the granite during late uplift in a young mountain belt.

Carbon-14 is a radioactive isotope of carbon (carbon-12 and carbon-13 are stable isotopes) with a half-life of 5,570 years. Carbon-14 is incorporated in all living material, for it is derived either directly or indirectly from its presence in atmospheric carbon dioxide. The moderately short half-life of carbon-14 makes it useful for dating biological materials that are more than a few hundred years old and less than 30,000 years old. It has been used to provide correlation of events within this time span, particularly those of the Pleistocene Epoch involving the Earth's most recent ice ages.

STUDY OF THE STRUCTURE OF THE EARTH

Geodesy. The scientific objective of geodesy is to determine the size and shape of the Earth. The practical role of geodesy is to provide a network of accurately surveyed points on the Earth's surface, the vertical elevations and geographic positions of which are precisely known and, in turn, may be incorporated in maps. When two geographic coordinates of a control point on the Earth's surface, its latitude and longitude, are known, as well as its elevation above sea level, the location of that point is known with an accuracy within the limits of error involved in the surveying processes. In mapping large areas, such as a whole state or country, the irregularities in the curvature of the Earth must be considered. A network of precisely surveyed control points provides a skeleton to which other surveys may be tied to provide progressively finer networks of more closely spaced points. The resulting networks of points have many uses, including anchor points or bench marks for surveys of highways and other civil features. A major use of control points is to provide reference points to which the contour lines and other features of topographic maps are tied. Most topographic maps are made using photogrammetric techniques and aerial photographs.

Establishing latitude and longitude

The Earth's figure is that of a surface called the geoid, which over the Earth is the average sea level at each location; under the continents the geoid is an imaginary continuation of sea level. The geoid is not a uniform spheroid, however, because of the existence of irregularities in the attraction of gravity from place to place on the Earth's surface. These irregularities of the geoid would bring about serious errors in the surveyed location of control points if astronomical methods, which involve use of the local horizon, were used solely in determining locations. Because of these irregularities, the reference surface used in geodesy is that of a regular mathematical surface, an ellipsoid of revolution that fits the geoid as closely as possible. This reference ellipsoid is below the geoid in some places and above it in others. Over the oceans, mean sea level defines the geoid surface, but over the land areas, the geoid is an imaginary sea-level surface.

The geoid and the spheroid

Today, perturbations in the motions of artificial satellites are used to define the global geoid and gravity pattern with a high degree of accuracy. Geodetic satellites are positioned at a height of 700–800 kilometres above the Earth. Simultaneous range observations from several laser stations fix the position of a satellite, and radar altimeters measure directly its height over the oceans. Results show that the geoid is irregular; in places its surface is up to 100 metres higher than the ideal reference ellipsoid and else-

where it is as much as 100 metres below it. The most likely explanation for this height variation is that the gravity (and density) anomalies are related to mantle convection and temperature differences at depth. An important observation that confirms this interpretation is that there is a close correlation between the gravity anomalies and the surface expression of the Earth's plate boundaries. This also strengthens the idea that the ultimate driving force of plate tectonics is a large-scale circulation of the mantle.

Geophysics. Geophysics pertains to studies of the Earth that involve the methods and principles of physics. The scope of geophysics touches on virtually all aspects of geology, ranging from considerations of the conditions in the Earth's deep interior, where temperatures of several thousands of degrees Celsius and pressures of millions of atmospheres prevail, to the Earth's exterior, including its atmosphere and hydrosphere.

The study of the Earth's interior provides a good example of the geophysicist's approach to problems. Direct observation is obviously impossible. Extensive knowledge of the Earth's interior has been derived from a variety of measurements, however, including seismic waves produced by quakes that travel through the Earth, measurements of the flow of heat from the Earth's interior into the outer crust, and by astronomical and other geologic considerations.

Geophysics may be divided into a number of overlapping branches in the following way: (1) study of the variations in the Earth's gravity field; (2) seismology, the study of the Earth's crust and interior by analysis of the transmission of elastic waves that are reflected or refracted; (3) the physics of the outer parts of the atmosphere, with particular attention to the radiation bombardment from the Sun and from outer space, including the influence of the Earth's magnetic field on radiation intercepted by the planet; (4) terrestrial electricity, which is the study of the storage and flow of electricity in the atmosphere and the solid Earth; (5) geomagnetism, the study of the source, configuration, and changes in the Earth's magnetic field and the study and interpretation of the remanent magnetism in rocks induced by the Earth's magnetic field when the rocks were formed (paleomagnetism); (6) the study of the Earth's thermal properties, including the temperature distribution of the Earth's interior and the variation in the transmission of heat from the interior to the surface; and (7) the convergence of several of the above-cited branches for the study of the large-scale structures of the Earth, such as rifts, continental margins, subduction zones, mid-oceanic ridges, thrusts, and continental sutures.

Geophysical techniques

The techniques of geophysics include measurement of the Earth's gravitational field using gravimeters on land and sea and artificial satellites in space (see above); measurement of its magnetic field with hand-held magnetometers or larger units towed behind research ships and aircraft; and seismographic measurement of subsurface structures using reflected and refracted elastic waves generated either by earthquakes or by artificial means (*e.g.*, underground nuclear explosions or ground vibrations produced with special pistons in large trucks). Other tools and techniques of geophysics are diverse. Some involve laboratory studies of rocks and other earth materials under high pressures and elevated temperatures. The transmission of elastic waves through the crust and interior of the Earth is strongly influenced by the behaviour of materials under the extreme conditions at depth; consequently, there is strong reason to attempt to simulate those conditions of elevated temperatures and pressures in the laboratory. At another extreme, data gathered by rockets and satellites yield much information about radiation flux in space and the magnetic effects of the Earth and other planetary bodies, as well as providing high precision in establishing locations in geodetic surveying, particularly over the oceans. Finally, it should be emphasized that the tools of geophysics are essentially mathematical and that most geophysical concepts are necessarily expounded mathematically.

Geophysics has major influence both as a field of pure science in which the objective is pursuit of knowledge for the sake of knowledge and as an applied science in which the objectives involve solution of problems of practical or commercial interest. Its principal commercial applications lie in the exploration for oil and natural gas and, to a lesser extent, in the search for metallic ore deposits. Geophysical methods also are used in certain geologic-engineering applications, as in determining the depth of alluvial fill that overlies bedrock, which is an important factor in the construction of highways and large buildings.

Much of the success of the plate tectonics theory has depended on the corroborative factual evidence provided by geophysical techniques. For example, seismology has demonstrated that the earthquake belts of the world demarcate the plate boundaries and that intermediate and deep seismic foci define the dip of subduction zones; the study of rock magnetism has defined the magnetic anomaly patterns of the oceans; and paleomagnetism has charted the drift of continents through geologic time. Seismic reflection profiling has revolutionized scientific ideas about the deep structure of the continents: major thrusts, such as the Wind River thrust in Wyoming and the Moine thrust in northwestern Scotland, can be seen on the profiles to extend from the surface to the Moho at about 35-kilometres depth; the Appalachian Mountains in the eastern United States must have been pushed at least 260 kilometres westward to their present position on a major thrust plane that now lies at about 15 kilometres depth; the thick crust of Tibet can be shown to consist of a stack of major thrust units; the shape and structure of continental margins against such oceans as the Atlantic and the Pacific are beautifully illustrated on the profiles; and the detailed structure of entire sedimentary basins can be studied in the search for oil reservoirs.

Structural geology. Structural geology deals with the geometric relationships of rocks and geologic features in general. The scope of structural geology is vast, extending over a scale of sizes ranging from submicroscopic lattice defects in crystals to mountain belts and plate boundaries.

Small-scale features. Structural features may be divided into two broad classes: the primary structures that were acquired in the genesis of a rock mass and the secondary structures that result from later deformation of the primary structures. Most layered rocks (sedimentary rocks, some lava flows, and pyroclastic deposits) were deposited initially as nearly horizontal layers. Rocks that were initially horizontal may be deformed later by folding and may be displaced along fractures. If displacement has occurred and the rocks on the two sides of the fracture have moved in opposite directions from each other, the fracture is termed a fault; if displacement has not occurred, the fracture is called a joint. It is clear that faults and joints are secondary structures—*i.e.*, their relative age is younger than the rocks that they intersect, but their age may be only slightly younger. Many joints in igneous rocks, for example, were produced by contraction when the rocks cooled. On the other hand, some fractures in rocks, including igneous rocks, are related to weathering processes and expansion associated with removal of overlying load. These will have been produced long after the rocks were formed. The faults and joints referred to above are brittle structures that form as discrete fractures within otherwise undeformed rocks in cool upper levels of the crust. In contrast, ductile structures result from permanent changes throughout a wide body of deformed rock at higher temperatures and pressures in deeper crustal levels. Such structures include folds and cleavage in slate belts, foliation in gneisses, and mineral lineation in metamorphic rocks.

Faults, joints, and mineral orientation

Large-scale features. Toward the other extreme are large-scale structural features, the study of which is termed tectonics or tectonophysics. These structures include mid-oceanic rifts; transform faults in the oceans; intracontinental rifts, as in the East African Rift System and on the Tibetan Highlands; wrench faults (*e.g.*, the San Andreas Fault in California) that may extend hundreds of kilometres; sedimentary basins (oil potential); thrusts, such as the Main Central thrust in the Himalayas, that measure more than 2,000 kilometres long; ophiolite complexes; passive continental margins, as around the Atlantic Ocean; active continental margins, as around the Pacific Ocean; trench systems at the mouth of subduction zones; granitic batholiths (*e.g.*, those in Sierra Nevada and Peru) that may be as long as 1,000 kilometres; complete sections of moun-

Tectonics

tain belts, such as the Andes, the Rockies, the Alps, the Himalayas, the Urals, and the Appalachians–Caledonians. Viewed as a whole, the study of these large-scale features encompasses the structural geology of plate tectonics.

Methodology. The methods of structural geology are diverse. At the smallest scale, lattice defects and dislocations in crystals can be studied in images enlarged several thousand times with transmission electron microscopes. Many structures can be examined microscopically, using the same general techniques employed in petrology, in which sections of rock mounted on glass slides are ground very thin and are then examined by transmitted light with polarizing microscopes. Of course, some structures can be studied in hand specimens, which were preferably oriented when collected in the field.

Field geology and the mapping of structural features

On a large scale, the techniques of field geology are employed. These include the preparation of geologic maps that show the areal distribution of geologic units selected for representation on the map. They also include the plotting of the orientation of such structural features as faults, joints, cleavage, small folds, and the attitude of beds with respect to three-dimensional space. A common objective is to interpret the structure at some depth below the surface. It is possible to infer with some degree of accuracy the structure beneath the surface by using information available at the surface. If geologic information from drill holes or mine openings is available, however, the configuration of rocks in the subsurface commonly may be interpreted with much greater assurance as compared with interpretations involving projection to depth based largely on information obtained at the surface. Vertical graphic sections are widely used to show the configuration of rocks beneath the surface. Balancing cross sections is an important technique in thrust belts. The lengths of individual thrust slices are added up and the total restored length is compared with the present length of the section and thus the percentage of shortening across the thrust belt can be calculated. In addition, contour maps that portray the elevation of particular layers with respect to sea level or some other datum are widely used, as are contour maps that represent thickness variations.

Strain analysis

Strain analysis is another important technique of structural geology. Strain is change in shape; for example, by measuring the elliptical shape of deformed ooliths or concretions that must originally have been circular, it is possible to make a quantitative analysis of the strain patterns in deformed sediments. Other useful kinds of specimens are deformed fossils, conglomerate pebbles, and vesicles. A long-term aim of such analysis is to determine the strain variations across entire segments of mountain belts. This information is expected to help geologists understand the mechanisms involved in the formation of such belts.

A combination of structural and geophysical methods are generally used to conduct field studies of the large-scale features mentioned above. Field work enables the mapping of the structures at the surface, and geophysical methods involving the study of seismic activity, magnetism, and gravity make possible the determination of the subsurface structures.

The processes that affect geologic structures rarely can be observed directly. The nature of the deforming forces and the manner in which the Earth's materials deform under stress can be studied experimentally and theoretically, however, thus providing insight into the forces of nature. One form of laboratory experimentation involves the deformation of small, cylindrical specimens of rocks under very high pressures. Other experimental methods include the use of scale models of folds and faults consisting of soft, layered materials, in which the objective is to simulate the behaviour of real strata that have undergone deformation on a larger scale over much longer time.

Some experiments measure the main physical variables that control rock deformation—namely, temperature, pressure, deformation rate, and the presence of fluids such as water. These variables are responsible for changing the rheology of rocks from rigid and brittle at or near the Earth's surface to weak and ductile at great depths. Thus experimental studies aim to define the conditions under which deformation occurs throughout the Earth's crust.

Volcanology. Volcanology is the science of volcanoes and deals with their structure, petrology, and origin. It is also concerned with the contribution of volcanoes to the rock structure of the Earth's crust, with their role as contributors to the atmosphere and hydrosphere and to the balance of chemical elements in the Earth's crust, and with the relationships of volcanoes to certain forms of metallic ore deposits.

Many of the problems of volcanology are closely related to those of the origin of oceans and continents. Most of the volcanoes of the world are aligned along or close to the major plate boundaries, in particular the mid-oceanic ridges and active continental margins (*e.g.*, the "ring of fire" around the Pacific Ocean). A few volcanoes occur within oceanic plates (*e.g.*, along the Hawaiian chain); these are interpreted as the tracks of plumes (ascending jets of partially molten mantle material) that formed when such a plate moved over hot spots fixed in the mantle.

One of the principal reasons for studying volcanoes and volcanic products is that the atmosphere and hydrosphere are believed to be largely derived from volcanic emanations, modified by biological processes. Much of the water present at the Earth's surface, which has aggregated mostly in the oceans but to a lesser extent in glaciers, streams, lakes, and groundwater, probably has emerged gradually from the Earth's interior by means of volcanoes, beginning very early in the Earth's history. The principal components of air—nitrogen and oxygen—probably have been derived through modification of ammonia and carbon dioxide emitted by volcanoes. Emissions of vapours and gases from volcanoes are an aspect of the degassing of the Earth's interior. Although the degassing processes that affect the Earth were probably much more vigorous when it was newly formed about 4,600,000,000 years ago, it is interesting to consider that the degassing processes are still at work. Their scale, however, is vastly reduced compared with their former intensity.

The study of volcanoes is dependent on a variety of techniques. The petrologic polarizing microscope is used for classifying lava types and for tracing their general mineralogical history. The X-ray fluorescence spectrometer provides a tool for making chemical analyses of rocks that are important for understanding the chemistry of a wide variety of volcanic products (*e.g.*, ashes, pumice, scoriae, and bombs) and of the magmas that give rise to them. Some lavas are enriched or depleted in certain isotopic ratios that can be determined with a mass spectrometer. Analyses of gases from volcanoes and of hot springs in volcanic regions provide information about the late stages of volcanic activity. These late stages are characterized by the emission of volatile materials, including sulfurous gases. Many commercially valuable ore deposits have formed through the influence of hydrothermal volcanic solutions.

Prediction of eruptions

Volcanoes may pose a serious hazard to human life and property, as borne out by the destruction wrought by the eruptions of Mount Vesuvius (AD 79), Krakatoa (1883), Mount Pelée (1902), and Mount Saint Helens (1980), to mention only a few. Because of this, much attention has been devoted to forecasting volcanic outbursts. In 1959 researchers monitored activity leading up to the eruption of Kilauea in Hawaii. Using seismographs, they detected swarms of earthquake tremors for several months prior to the eruption, noting a sharp increase in the number and intensity of small quakes shortly before the outpouring of lava. Tracking such tremors, which are generated by the upward movement of magma from the asthenosphere, has proved to be an effective means of determining the onset of eruptions and is now widely used for prediction purposes. Some volcanoes inflate when rising molten rock fills their magma chambers, and in such cases tiltmeters can be employed to detect a change in angle of the slope before eruption. Other methods of predicting violent volcanic activity involve the use of laser beams to check for changes in slope, temperature monitors, gas detectors, and instruments sensitive to variations in magnetic and gravity fields. Permanent volcano observatories have been established at some of the world's most active sites (*e.g.*, Kilauea, Mount Etna, and Mount Saint Helens) to ensure early warning.

STUDY OF SURFACE FEATURES AND PROCESSES

Geomorphology. Geomorphology is literally the study of the form or shape of the Earth, but it deals principally with the topographical features of the Earth's surface. It is concerned with the classification, description, and origin of landforms. The configuration of the Earth's surface reflects to some degree virtually all of the processes that take place at or close to the surface as well as those that occur deep in the crust. The intricate details of the shape of a mountain range, for example, result more or less directly from the processes of erosion that progressively remove material from the range. The spectrum of erosive processes includes weathering and soil-forming processes and transportation of materials by running water, wind action, and mass movement. Glacial processes have been particularly influential in many mountainous regions. These processes are destructional in the sense that they modify and gradually destroy the previous form of the range. Also important in governing the external shape of the range are the constructional processes that are responsible for uplift of the mass of rock from which the range has been sculptured. A volcanic cone, for example, may be created by the successive outpouring of lava, perhaps coupled with intermittent ejection of volcanic ash and tuff. If the cone has been built up rapidly, so that there has been relatively little time for erosive processes to modify its form, its shape is governed chiefly by the constructional processes involved in the outpouring of volcanic material. But the forces of erosion begin to modify the shape of a volcanic landform almost immediately and continue indefinitely. Thus, at no time can its shape be regarded as purely constructional or purely destructional, for its shape is necessarily a consequence of the interplay of these two major classes of processes.

Destructional and constructional geomorphic processes

Investigating the processes that influence landforms is an important aspect of geomorphology. These processes include the weathering caused by the action of solutions of atmospheric carbon dioxide and oxygen in water on exposed rocks; the activity of streams and lakes; the transport and deposition of dust and sand by wind; the movement of material through downhill creep of soil and rock and by landslides and mudflows; and shoreline processes that involve the mechanics and effects of waves and currents. Study of these different types of processes forms subdisciplines that exist more or less in their own right.

Glacial geology. Glacial geology can be regarded as a branch of geomorphology, though it is such a large area of research that it stands as a distinct subdiscipline within the geologic sciences. Glacial geology is concerned with the properties of glaciers themselves as well as with the effects of glaciers as agents of both erosion and deposition. Glaciers are accumulations of snow transformed into solid ice. Important questions of glacial geology concern the climatic controls that influence the occurrence of glaciers, the processes by which snow is transformed into ice, and the mechanism of the flow of ice within glaciers. Other important questions involve the manner in which glaciers serve as erosive agents, not only in mountainous regions but also over large regions where great continental glaciers now extend or once existed. Much of the topography of the northern part of North America and Eurasia, for example, has been strongly influenced by glaciers. In places, bedrock has been scoured of most surficial debris. Elsewhere, deposits of glacial till mantle much of the area. Other extensive deposits include unconsolidated sediments deposited in former lakes that existed temporarily as a result of dams created by glacial ice or by glacial deposits. Many presently existing lakes are of glacial origin as, for example, the Great Lakes.

Glacial features and deposits

Research in glacial geology is conducted with a variety of tools. Investigators use, for example, radar techniques to determine the thickness of glaciers. In order to calculate the progressive advance or retreat of glacial masses, they ascertain the age of organic materials associated with glacial moraines by means of isotopic analyses.

Other branches of the geologic sciences are closely linked with glacial geology. In glaciated regions the problems of hydrology and hydrogeology are strongly influenced by the presence of glacial deposits. Furthermore, the suitability of glacial deposits as sites for buildings, roads, and other man-made features is influenced by the mechanical properties of the deposits and by soils formed on them.

EARTH HISTORY

Historical geology and stratigraphy. One of the major objectives of geology is to establish the history of the Earth from its inception to the present. The most important evidence from which geologic history can be inferred is provided by the geometric relationships of rocks with respect to each other, particularly layered rocks, or strata, the relative ages of which may be determined by applying simple principles. One of the major principles of stratigraphy is that within a sequence of layers of sedimentary rock, the oldest layer is at the base and that the layers are progressively younger with ascending order in the sequence. This is termed the law of superposition and is one of the great general principles of geology. Ordinarily, beds of sedimentary rocks are deposited more or less horizontally. In some regions sedimentary strata have remained more or less horizontal long after they were deposited. Some of these sedimentary rocks were deposited in shallow seas that once extended over large areas of the present continents. In many places sedimentary rocks lie much above sea level, reflecting vertical shift of the crust relative to sea level. In regions where the rocks have been strongly deformed through folding or faulting, the original attitudes of strata may be greatly altered, and sequences of strata that were once essentially horizontal may now be steeply inclined or overturned.

The law of superposition as a guiding principle

Prior to the development of radiometric methods of dating rocks, the ages of rocks and other geologic features could not be expressed quantitatively, or as numbers of years, but instead were expressed solely in terms of relative ages, in which the age of a particular geologic feature could be expressed as relatively younger or older than other geologic features. The ages of different sequences of strata, for example, can be compared with each other in this manner, and their relative ages with respect to faults, igneous intrusions, and other features that exhibit crosscutting relationships can be established. Given such a network of relative ages, a chronology of events has been gradually established in which the relative time of origin of various geologic features is known. This is the main thread of historical geology—an ordered sequence of geologic events whose occurrence and relative ages have been inferred from evidence preserved in the rocks. In turn, the development of radiometric dating methods has permitted numerical estimates of age to be incorporated in the scale of geologic time.

The development of the mass spectrometer has provided researchers with a means of calculating quantitative ages for rocks throughout the whole of the geologic record. With the aid of various radiometric methods involving mass spectrometric analysis, researchers have found it possible to determine how long ago a particular sediment was deposited, when an igneous rock crystallized or when a metamorphic rock recrystallized, and even the time at which rocks in a mountain belt cooled or underwent uplift. Radiometric dating also helped geochronologists discover the vast span of geologic time. The radiometric dating of meteorites revealed that the Earth, like other bodies of the solar system, is about 4,600,000,000 years old and that the oldest rocks so far discovered formed roughly 3,800,000,000 years ago. It has been established that the Precambrian time occupies seven-eighths of geologic time, but the era is still poorly understood in comparison with the Phanerozoic Eon—the span of time extending from about the beginning of the Cambrian Period to the Holocene Epoch during which complex life forms are known to have existed. The success of dating Phanerozoic time with some degree of precision has depended on the interlinking of radiometric ages with biostratigraphy, which is the correlation of strata with fossils.

Importance of radiometric dating

Paleontology. The geologic time scale is based principally on the relative ages of sequences of sedimentary strata. Establishing the ages of strata within a region, as well as the ages of strata in other regions and on different continents, involves stratigraphic correlation from place to

place. Although correlation of strata over modest distances often can be accomplished by tracing particular beds from place to place, correlation over long distances and over the oceans almost invariably involves comparison of fossils. With rare exceptions, fossils occur only in sedimentary strata. Paleontology, which is the science of ancient life and deals with fossils, is mutually interdependent with stratigraphy and with historical geology. Paleontology also may be considered to be a branch of biology.

Organic evolution and index fossils

Organic evolution is the essential principle involved in the use of fossils for stratigraphic correlation. It incorporates progressive irreversible changes in the succession of organisms through time. A small proportion of types of organisms has undergone little or no apparent change over long intervals of geologic time, but most organisms have progressively changed, and earlier forms have become extinct and, in turn, have been succeeded by more modern forms. Organisms preserved as fossils that lived over a relatively short span of geologic time and that were geographically widespread are particularly useful for stratigraphic correlation. These fossils are indexes of relative geologic age and may be termed index fossils.

Fossils play another major role in geology because they serve as indicators of ancient environments. Specialists called paleoecologists seek to determine the environmental conditions under which a fossil organism lived and the physical and biological constraints on those conditions. Did the organism live in the seas, lakes, or bogs? In what type of biological community did it live? What was its food chain? In short, what ecological niche did the organism occupy? Because oil and natural gas only accumulate in restricted environments, paleoecology can offer useful information for fossil fuel exploration.

Invertebrate paleontology. One of the major branches of paleontology is invertebrate paleontology, which is principally concerned with fossil marine invertebrate animals large enough to be seen with little or no magnification. The number of invertebrate fossil forms is large and includes brachiopods, pelecypods, cephalopods, gastropods, corals and other coelenterates (*e.g.*, jellyfish), bryozoans, sponges, various arthropods (invertebrates with limbs—*e.g.*, insects), including trilobites, echinoderms, and many other forms, some of which have no living counterparts. The invertebrates that are used as index fossils generally possess hard parts, a characteristic that has fostered their preservation as fossils. The hard parts preserved include the calcareous or chitinous shells of the brachiopods, cephalopods, pelecypods, and gastropods, the jointed exoskeletons of such arthropods as the trilobites, and the calcareous skeletons of frame-building corals and bryozoans. The vast variety of organisms lacking hard parts are poorly represented in the geologic record; however, they have sometimes been found to occur as impressions or carbonized films in finely laminated sediments.

Vertebrate paleontology. Vertebrate paleontology is concerned with fossils of the vertebrates: fish, amphibians, reptiles, birds, and mammals. Although vertebrate paleontology has close ties with stratigraphy, vertebrate fossils usually have not been extensively used as index fossils for stratigraphic correlation, vertebrates generally being much larger than invertebrate fossils and consequently rarer. Fossil mammals, however, have been widely used as index fossils for correlating certain nonmarine strata deposited during the Tertiary Period (from 66,400,000 to 1,600,000 years ago).

Sizes of microfossils

Micropaleontology. Micropaleontology involves the study of organisms so small that they can be observed only with the aid of a microscope. The size range of microscopic fossils, however, is immense. In most cases, the term micropaleontology connotes that aspect of paleontology devoted to the Ostracoda, a subclass of crustaceans that are generally less than one millimetre in length; Radiolaria, marine (typically planktonic) protozoans whose remains are common in deep ocean-floor sediments; and Foraminifera, marine protozoans that range in size from about 10 centimetres to a fraction of a millimetre.

Generally speaking, micropaleontology involves successive ranges of sizes of microscopic fossils down to organisms that must be magnified hundreds of times or more for viewing. The study of ultrasmall fossils is perhaps the fastest growing segment of contemporary paleontology and is dependent on modern laboratory instruments, including electron microscopes. It is an important aspect of oil and natural-gas exploration. Microfossils, which are flushed up boreholes in the drilling mud, can be analyzed to determine the depositional environment of the underlying sedimentary rocks and their age. This information enables geologists to evaluate the reservoir potential of the rock (*i.e.*, its capacity for holding gas or oil) and its depth. Ostracods and foraminiferans occur in such abundance and in so many varieties and shapes that they provide the basis for a detailed classification and time division of Mesozoic and Cenozoic sediments in which oil may occur.

Filamentous and spheroidal microfossils are important in many Precambrian sediments such as chert. They occur in rocks as old as 3,500,000,000 years and are thus an important testimony of early life on Earth.

Paleobotany. Paleobotany is the study of fossil plants. The oldest widely occurring fossils are various forms of calcareous algae that apparently lived in shallow seas, although some may have lived in freshwater. Their variety is so profuse that their study forms an important branch of paleobotany. Other forms of fossil plants consist of land plants or of plants that lived in swamp forests, standing in water that was fresh or may have been brackish, such as the coal-forming swamps of the Late Carboniferous Period (from 320,000,000 to 286,000,000 years ago).

Palynology. Palynology deals with plant spores and pollen that are both ancient and modern and is a branch of paleobotany. It plays an important role in the investigation of ancient climates, particularly through studies of deposits formed during glacial and interglacial stages. Study of a sequence of spore- or pollen-bearing beds may reveal successive climatic changes, as indicated by changes in types of spores and pollen derived from different vegetative complexes. Spores and pollen are borne by the wind and spread over large areas. Furthermore, they tend to be resistant to decay and thus may be preserved in sediments under adverse conditions.

Astrogeology. Astrogeology is concerned with the geology of the solid bodies in the solar system, such as the asteroids and the planets and their moons. Research in this field helps scientists to better understand the evolution of the Earth in comparison with that of its neighbours in the solar system. This subject was once the domain of astronomers, but the advent of spacecraft has made it accessible to geologists, geophysicists, and geochemists. The success of this field of study has depended largely on the development of advanced instrumentation.

Studies of the Moon

The U.S. Apollo program enabled humans to land on the Moon several times since 1969. Rocks were collected, geophysical experiments were set up on the lunar surface, and geophysical measurements were made from spacecraft. The Soyuz program of the Soviet Union also collected much geophysical data from orbiting spacecraft. The mineralogy, petrology, geochemistry, and geochronology of lunar rocks were studied in detail, and this research made it possible to work out the geochemical evolution of the Moon. The various manned and unmanned missions to the Moon resulted in many other accomplishments: for example, a lunar stratigraphy was constructed; geologic maps at a scale of 1:1,000,000 were prepared; the structure of the maria, rilles, and craters was studied; gravity profiles across the dense, lava-filled maria were produced; the distribution of heat-producing radioactive elements, such as uranium and thorium, was mapped with gamma-ray spectrometers; the Moon's internal structure was determined on the basis of seismographic records of moonquakes; the heat flow from the interior was measured; and the day and night temperatures at the surface were recorded.

Exploration of the inner planets

From the late 1960s to the early 1990s, unmanned spacecraft were sent to the neighbouring planets by American and Soviet scientists. Several of these probes were soft-landed on Mars and Venus. Soil scoops from the Martian surface have been chemically analyzed by an on-board X-ray fluorescence spectrometer. The radioactivity of the surface materials of both Mars and Venus have been studied with a gamma-ray detector, the isotopic composition

of their atmospheres analyzed with a mass spectrometer, and their magnetic fields measured. Relief and geologic maps of Mars have been made from high-resolution photographs and topographical maps of Venus compiled from radar data transmitted by orbiting spacecraft. Photographs of Mars and Mercury show that their surfaces are studded with many meteorite craters similar to those on the Moon. Detailed studies have been made of the craters, volcanic landforms, lava flows, and rift valleys on Mars, and a simplified geologic-thermal history has been constructed for the planet.

By the mid-1980s, the United States had sent interplanetary probes past Jupiter, Saturn, and Uranus. The craft transmitted data and high-resolution photographs of these outer planetary systems, including their rings and satellites.

This research has given increased impetus to the study of tektites, meteorites, and meteorite craters on Earth. The mineralogy, geochemistry, and isotopic age of meteorites and tektites have been studied in detail. Meteorites are very old and probably originated in the asteroid belt between Mars and Jupiter, while tektites are very young and most likely formed from material ejected from terrestrial meteorite craters. Many comparative studies have been made of the development and shapes of meteorite craters on Earth, the Moon, Mars, and Mercury. Space exploration has given birth to a new science—the geology of the solar system. The Earth can now be understood within the framework of planetary evolution.

PRACTICAL APPLICATIONS

Exploration for energy and mineral sources. Over the past century, industries have developed rapidly, populations have grown dramatically, and standards of living have improved, resulting in an ever-growing demand for energy and mineral resources. Geologists and geophysicists have led the exploration for fossil fuels (coal, oil, natural gas, etc.) and concentrations of geothermal energy, for which applications have grown in recent years. They also have played a major role in locating deposits of commercially valuable minerals.

Coal. The Industrial Revolution of the late 18th and 19th centuries was fueled by coal. Though it has been supplanted by oil and natural gas as the primary source of energy in most modern industrial nations, coal nonetheless remains an important fuel.

The U.S. Geological Survey has estimated that only about 2 percent of the world's minable coal has so far been exploited; known reserves should last for at least 300 to 400 years. Moreover, new coal basins continue to be found, as, for example, the lignite basin discovered in the mid-1980s in Rājasthān in northwestern India.

Coal-exploration geologists have found that coal was formed in two different tectonic settings: (1) swampy marine deltas on stable continental margins, and (2) swampy freshwater lakes in graben (long, narrow troughs between two parallel normal faults) on continental crust. Knowing this and the types of sedimentary rock formations that commonly include coal, geologists can quite readily locate coal-bearing areas. Their main concern, therefore, is the quality of the coal and the thickness of the coal bed or seam. (A coal seam must be at least 61 centimetres thick to be mined profitably.) Such information can be derived from samples obtained by drilling into the rock formation in which the coal occurs.

Locating coal-bearing areas

Oil and natural gas. During the last half of the 20th century, the consumption of petroleum products increased sharply. This has led to a depletion of many existing oil fields, notably in the United States, and intensive efforts to find new deposits.

Crude oil and natural gas in commercial quantities are generally found in sedimentary rocks along rifted continental margins and in intracontinental basins. Such environments exhibit the particular combination of geologic conditions and rock types and structures conducive to the formation and accumulation of liquid and gaseous hydrocarbons. They contain suitable source rocks (organically rich sedimentary rocks such as black shale), reservoir rocks (those of high porosity and permeability capable of holding the oil and gas that migrate into them), and overlying impermeable rocks that prevent the further upward movement of the fluids. These so-called cap rocks form petroleum traps, which may be either structural or stratigraphic depending on whether they were produced by crustal deformation or original sedimentation patterns.

Techniques of petroleum exploration

Petroleum geologists concentrate their search for oil deposits in such geologic settings, mapping both the surface and subsurface features of a promising area in great detail. Geologic surface maps show subcropping sedimentary rocks and features associated with structural traps such as ridges formed by anticlines during the early stages of folding and lineations produced by fault ruptures. Maps of this kind may be based on direct observation or may be constructed with photographs taken from aircraft and Earth-orbiting satellites, particularly of terrain in remote areas. Subsurface maps reveal possible hidden underground structures and lateral variations in sedimentary rock bodies that might form a petroleum trap. The presence of such features can be detected by various means, including gravity measurements, seismic methods, and the analysis of borehole samples from exploratory drilling. (For a description of these techniques, see EXPLORATION: *Exploration of the Earth's surface and interior.*)

Another method used by petroleum geologists in exploratory areas involves the sampling of surface waters from swamps, streams, or lakes. The water samples are analyzed for traces of hydrocarbons, the presence of which would indicate seepage from a subsurface petroleum trap. This geochemical technique, along with seismic profiling, is often used to search for offshore petroleum accumulations.

Once an oil deposit has actually been located and well drilling is under way, petroleum geologists can determine from core samples the depth and thickness of the reservoir rock as well as its porosity and permeability. Such information enables them to estimate the quantity of the oil present and the ease with which it can be recovered.

Although only about 15 percent of the world's oil has been exploited, petroleum geologists estimate that at the present rate of demand the supply of recoverable oil will last no more than 100 years. Owing to this rapid depletion of conventional oil sources, economic geologists have explored oil shales and tar sands as potential supplementary petroleum resources. Extracting oil from these substances is, however, very expensive and involves possible environmental problems. But both are abundant, and advances in recovery technology may yet make them attractive alternative energy resources.

Oil shales and tar sands as possible alternative energy resources

Geothermal energy. Another alternate energy resource is the heat from the Earth's interior. The surface expression of this energy is manifested in volcanoes, fumaroles, steam geysers, hot springs, and boiling mud pools. Global heat-flow maps constructed from geophysical data show that the zones of highest heat flow occur along the active plate boundaries. There is, in effect, a close association between geothermal energy sources and volcanically active regions.

A variety of applications have been developed for geothermal energy. For example, public buildings, residential dwellings, and greenhouses in such areas as Reykjavík, Ice., are heated with water pumped from hot springs and geothermal wells. Hot water from such sources also is used for heating soil to increase crop production (*e.g.*, in Oregon) and for seasoning lumber (*e.g.*, in parts of New Zealand). The most significant application of geothermal energy, however, is the generation of electricity. The first geothermal power station began operation in Larderello, Italy, in the early 1900s. Since then similar facilities have been built in various countries, including Iceland, Japan, Mexico, New Zealand, Turkey, and the United States. In most cases turbines are driven with steam separated from superheated water tapped from underground geothermal reservoirs and geysers.

Mineral deposits. As was mentioned above, the distribution of commercially significant mineral deposits, the economic factors associated with their recovery, and the estimates of available reserves constitute the basic concerns of economic geologists. Because continued industrial development is heavily dependent on mineral resources, their work is crucial to modern society.

Mineral deposits in particular types of plate-tectonic environments

It has long been known that certain periods of Earth history were especially favourable for the concentration of specific types of minerals. Copper, zinc, nickel, and gold are important in Archean rocks; magnetite and hematite are concentrated in early Proterozoic banded-iron formations; and there are economic Proterozoic uranium reserves in conglomerates. These mineral deposits and a variety of others that developed throughout the Phanerozoic Eon can be related to specific types of plate-tectonic environments. Among the latter are copper, lead, and zinc in intracontinental rifts. An interesting discovery has been the remarkable concentrations of gold, iron, zinc, and copper in brine pools and sulfide-rich muds in the Red Sea and in the Salton Sea in southern California. In many countries copper, nickel, and chromium deposits occur in ophiolite complexes obducted onto the continents from the ocean floor; porphyry copper and molybdenum deposits are found in association with granodioritic intrusions; and tungsten and tin deposits occur in many granites. The correlation of these associations and distributions with periods of Earth history, on the one hand, and plate-tectonic settings, on the other, have enabled regional metallogenetic provinces to be defined, which have proved helpful in the search for ore deposits.

During the 20th century the exploitation of mineral deposits has been so intense that serious depletion of many resources is predicted. Mercury reserves, for example, are particularly low. To deal with this problem, it has become necessary to mine deposits having smaller and smaller workable grades, a trend well illustrated by the copper mining industry, which now extracts copper from rocks with grades as low as 0.2 percent.

Investigators have discovered a major potential metallic source on the deep ocean floor, where there are large concentrations of manganese-rich nodules along with minor amounts of copper, nickel, and cobalt. Such concentrations are especially abundant in three sections of the Pacific Ocean—the area near Hawaii, that northeast of New Zealand, and that west of Central America.

Earthquake prediction and control. No natural event is as destructive over so large an area in so short a time as an earthquake. Throughout the centuries earthquakes have been responsible not only for millions of deaths but also for tremendous damage to property and the natural landscape. If major earthquakes could be predicted, it would be possible to evacuate population centres and take other measures that could minimize the loss of life and perhaps reduce damage to property as well. For this reason earthquake prediction has become a major concern of seismologists in the United States, the Soviet Union, Japan, and China.

World seismicity patterns show that earthquakes tend to occur along active plate boundaries where there is subduction (Japan) or strike-slip motion (California) and along strike-slip faults (as in China, where they are the result of the northward migration of India into Asia). Investigators agree that much more has to be learned about the physical properties of rocks in fault zones before they are able to make use of changes in these properties to predict earthquakes. Recent research has suggested that rocks may become strained shortly before an earthquake and affect such observable properties of the Earth's crust as seismic wave velocity and radon concentration. Leveling surveys and tiltmeter measurements have revealed that deformation in the fault zone just prior to an earthquake may cause changes in ground level and, in certain cases, variations in groundwater level. Also, some investigators have reported changes in the electric resistivity and remanent magnetization of rocks as precursory phenomena.

Possible precursory phenomena

Since the San Francisco earthquake of 1906, seismic activity along the nearby San Andreas Fault has been closely monitored. It has been observed that numerous semicontinuous microearthquakes have occurred along some sections of the fault. These small quakes seem to release built-up strain and thus prevent large earthquakes. By contrast, intervening sections of the fault are apparently locked and thus manifest no microshocks. Consequently, seismic strain accumulating in these locked sections is expected to be released one day in a major quake.

Seismological research includes the study of earthquakes caused by human activities, such as impounding water behind high dams, injecting fluids into deep wells, excavating mines, and detonating underground nuclear explosions. In all of these cases except for deep mining, seismologists have found that the induction mechanism most likely involves the release of elastic strain, just as with earthquakes of tectonic origin. Studies of artificially induced quakes suggest that one possible method of controlling natural earthquakes is to inject fluids into fault zones so as to release strain energy.

Artificially induced quakes

Seismologists have done much to explain the characteristics of ground motions recorded in earthquakes. Such information is required to predict ground motions in future earthquakes, thereby enabling engineers to design earthquake-resistant structures. The largest percentage of the deaths and property damage that result from an earthquake is attributable to the collapse of buildings, bridges, and other man-made structures during the violent shaking of the ground. An effective way of reducing the destructiveness of earthquakes, therefore, is to build structures capable of withstanding intense ground motions.

Other areas of application. The fields of engineering, environmental, and urban geology are broadly concerned with applying the findings of geologic studies to construction engineering and to problems of land use. The location of a bridge, for example, involves geologic considerations in selecting sites for the supporting piers. The strength of geologic materials such as rock or compacted clay that occur at the sites of the piers should be adequate to support the load placed on them. Engineering geology is concerned with the engineering properties of geologic materials, including their strength, permeability, and compactability, and with the influence of these properties on the selection of locations for buildings, roads and railroads, bridges, dams, and other major civil features.

Urban and environmental geology

Urban geology involves the application of engineering geology and other fields of geology to environmental problems in urban areas. Environmental geology is generally concerned with those aspects of geology that touch on the human environment. Environmental and urban geology deal in large measure with those aspects of geology that directly influence land use. These include the stability of sites for buildings and other civil features, sources of water supply, contamination of waters by sewage and chemical pollutants, selection of sites for burial of refuse so as to minimize pollution by seepage, and locating the source of geologic building materials, including sand, gravel, and crushed rock.

(J.W.Ha./B.F.W.)

Hydrologic sciences

Scope of the hydrologic sciences

The hydrologic sciences deal with the study of the waters of the Earth. In its widest sense hydrology encompasses the study of the occurrence, the movement, and the physical and chemical characteristics of water in all its forms within the Earth's hydrosphere. In practice hydrologists usually restrict their studies to waters close to the land surface of the Earth. Water in the atmosphere is usually studied as part of meteorology (see below *Atmospheric sciences*). Water in the oceans and seas is studied within the science of oceanography; water in lakes and inland seas within limnology; and ice on the land surface within glaciology. Clearly there is some overlap between these major scientific disciplines; both hydrologists and meteorologists, for example, have contributed to the study of water movement in the lower boundary layers of the atmosphere. All are linked by the fundamental concept of the hydrologic cycle, according to which the waters of the sea are evaporated, are subsequently condensed within the atmosphere, fall to the Earth as precipitation, and finally flow in the rivers back to the sea.

Water is the most abundant substance on Earth and is the principal constituent of all living things. Water in the atmosphere plays a major role in maintaining a habitable environment for human life. The occurrence of surface waters has played a significant role in the rise and decline of the major civilizations in world history. In many societies the importance of water to humankind is reflected in

the legal and political structures. At the present time rising populations and improving living standards are placing increasing pressures on available water resources. There is, in general, no shortage of water on the Earth's land surface, but the areas of surplus water are often located far from major centres of population. Moreover, in many cases these centres prove to be sources of water pollution. Thus, the availability and quality of water are becoming an ever-increasing constraint on human activities, notwithstanding the great technological advances that have been made in the control of surface waters.

STUDY OF THE WATERS CLOSE TO THE LAND SURFACE

Hydrology and its component disciplines

Hydrology deals with that part of the hydrologic cycle from the arrival of water at the land surface as precipitation to its eventual loss from the land either by evaporation or transpiration back to the atmosphere or by surface and subsurface flow to the sea. It is thus primarily concerned with waters close to the land surface. It includes various component disciplines of a more specialized nature. Hydraulics is concerned with the mechanics and dynamics of water in its liquid state. Hydrography is the description and mapping of the bodies of water of the Earth's surface (including the oceans), with a particular concern for navigation charts. Hydrometry involves measurements of surface water, particularly precipitation and streamflow. Hydrometeorology focuses on water in the lower boundary layer of the atmosphere. Groundwater hydrology and hydrogeology have to do with subsurface water in the saturated zone, while soil water physics involves the study of subsurface water in the unsaturated zone. Engineering hydrology is concerned with the design of man-made structures that control the flow and use of water.

Concept of water balance

Underlying all the hydrologic sciences is the concept of water balance, an expression of the hydrologic cycle for an area of the land surface in terms of conservation of mass. In a simple form the water balance may be expressed as

$$S = P - Q - E - G,$$

where S is the change of water storage in the area over a given time period, P is the precipitation input during that time period, Q is the stream discharge from the area, E is the total of evaporation and transpiration to the atmosphere from the area, and G is the subsurface outflow. Most hydrologic studies are concerned with evaluating one or more terms of the water balance equation. Because of the difficulties in quantifying the movement of water across the boundaries of an area under study, the water balance equation is most easily applied to an area draining to a particular measurement point on a stream channel. This area is called a catchment (or sometimes a watershed in the United States). The line separating adjacent catchments is known as a topographical divide, or simply a divide. The following sections describe the study of the different elements of the catchment water balance and the way in which they affect the response of catchments over time under different climatic regimes.

Evaluation of the catchment water balance. *Precipitation.* Precipitation results from the condensation of water from the atmosphere as air is cooled to the dew point, the temperature at which the air becomes saturated with respect to water vapour. The cooling process is usually initiated by uplift of the air, which may result from a number of causes, including convection, orographic effects over mountain ranges, or frontal effects at the boundaries of air masses of different characteristics. Condensation within the atmosphere requires the presence of condensation nuclei to initiate droplet formation. Some of the condensate may be carried considerable distances as cloud before being released as rain or snow, depending on the local temperatures. Some precipitation in the form of dew or fog results from condensation at or near the land surface. In some areas, such as the coastal northwest of the United States, dew and fog drip can contribute significantly to the water balance. The formation of hail requires a sequence of condensation and freezing episodes, resulting from successive periods of uplift. Hailstones usually show a pattern of concentric rings of ice as a result.

Direct measurement of rainfall

Direct measurements of precipitation are made by a variety of gauges, all of which consist of some form of funnel that directs the infalling water to some storage container. Storage gauges simply store the incident precipitation, and the accumulated water is usually measured on a daily, weekly, or monthly basis. Recording gauges allow rates of precipitation to be determined.

Rainfall volumes are usually converted to units of depth—volume per unit area. Measurements obtained from different types of rain gauges are not directly comparable because of varying exposure, wind, and splash effects. The most accurate type of gauge is the ground-level gauge, in which the orifice of the gauge is placed level with the ground surface and surrounded by an antisplash grid. Rain gauge catches decrease as the orifice is raised above the ground, particularly in areas subject to high winds. In areas of significant snowfall, however, it may be necessary to raise the rain gauge so that its orifice is clear of the snow surface. Various shields for the gauge orifice have been tried in an effort to offset wind effects. Wind effects are greater for snow than for rain and for small drops or light rainfall than for large drops.

Indirect measurements of rainfall

An impression of the spatial distribution of precipitation intensity can be achieved through indirect measurements of precipitation, in particular radar scattering. The relationship between rainfall intensity and measured radar signals depends on various factors, including the type of precipitation and the distribution of drop size. Radar measurements are often used in conjunction with rain gauges to allow on-line calibration in converting the radar signal to precipitation amounts. The radar measurements are, however, at a much larger spatial scale. Resolution of five to 10 square kilometres is common for operational systems. Even so, this provides a much better picture of the spatial patterns of precipitation over large catchment areas than has been previously possible. The use of satellite remote sensing to determine rainfall volumes is still in its early stages, but the technique appears likely to prove useful for estimating amounts of precipitation in remote areas.

Measuring snowfall

The measurement of inputs of snow to the catchment water balance is also a difficult problem. The most basic technique involves the snow course, a series of stakes to measure snow depths. Snowfalls can, however, vary greatly in density, depending primarily on the temperature history of snow formation. Accumulated snow changes its density over time prior to melting. Snow density can be measured by weighing a sample of known volume taken in a standard metal cylinder. Other techniques for measuring snowfall include the use of snow pillows, which record the changing weight of snow lying above them, or the use of rain gauges fitted with heating elements, which melt the snow as it falls. These techniques are subject to wind effects, both during a storm event and between events because of redistribution of snow by the wind.

Summary statistics on precipitation are usually produced on the basis of daily, monthly, and annual amounts falling at a given location or over a catchment area. The frequency at which a rainfall of a certain volume occurs within a certain period is also important to hydrologic analysis. The assessment of this frequency, or the recurrence interval of the rainfall from the sample of available data, is a statistical problem generally involving the assumption of a particular probability distribution to represent the characteristics of rainfalls. Such analyses must assume that this distribution is not changing over time, even though it has been shown that in some areas of the world climatic change may cause rainfall statistics to vary. It has long been speculated that rainfalls may exhibit cyclic patterns over long periods of time, and considerable effort has been expended in searching for such cycles. In some areas the annual seasonal cycle is of paramount importance, but demonstrations of longer periodicities have not proved of general applicability.

Patterns of rainfall intensity and duration are of great importance to the hydrologist in predicting catchment discharges and water availability and in dealing with floods, droughts, land drainage, and soil erosion. Rainfalls vary both within and between rainstorms, sometimes dramatically, depending on the type and scale of the storm and its

velocity of movement. Within a storm, the average intensity tends to decrease with an increase in the storm area.

On a larger scale, seasonal variations in rainfall vary with climate. Humid temperate areas tend to have rainfalls that are fairly evenly distributed throughout the year; Mediterranean areas have a winter peak with low summer rainfalls; savanna areas have a double peak in rainfall; and equatorial areas again have a relatively even distribution of rainfall over the course of the year. Average annual rainfalls also vary considerably. The minimum recorded long-term average is 0.76 millimetre at Arica, Chile; the maximum 11,897.36 millimetres at Tutunendo, Colom. The maximum recorded rainfall intensities are 38 millimetres in one minute (Barot, Guadeloupe, 1970); 1,870 millimetres in a single day (Cilaos, Réunion, 1952); and 26,461 millimetres in one year (Cherrapunji, India, 1861).

Interception. When precipitation reaches the surface in vegetated areas, a certain percentage of it is retained on or intercepted by the vegetation. Rainfall that is not intercepted is referred to as throughfall. Water that reaches the ground via the trunks and stems of the vegetation is called stemflow. The interception storage capacities of the vegetation vary with the type and structure of the vegetation and with meteorologic factors. Measurements have shown that up to eight millimetres of rainfall can be intercepted by some vegetation canopies. The intercepted water is evaporated back into the atmosphere at rates determined by the prevailing meteorologic conditions and the nature of the vegetation. In humid temperate areas evaporation of intercepted water can be an important component of the water balance. Forest areas have been shown to have greater interception losses than adjacent grassland areas. This is due to the greater aerodynamic roughness of the forest canopy, resulting in a much more efficient transfer of water vapour away from the surface.

Infiltration. When water from a rainstorm or a period of snowmelt reaches the ground, some or all of it will infiltrate the soil. The rate of infiltration depends on the intensity of the input, the initial moisture condition of the surface soil layer, and the hydraulic characteristics of the soil. Small-scale effects such as the presence of a surface seal of low permeability (due to the rearrangement of surface soil particles by rain splash) or the presence of large channels and cracks in the surface soil may be important in controlling infiltration rates. Water in excess of the infiltration capacity of the soil will flow overland as surface runoff once the minor undulations in the surface (the depression storage) have been filled. Such runoff occurs most frequently on bare soils and in areas subject to high rainfall intensities. In many environments rainfall intensities rarely exceed the infiltration capacities of vegetated soil surfaces. The occurrence of surface runoff is then more likely to be generated by rainfall on completely saturated soil.

Factors affecting the infiltration rate

Evapotranspiration. Rates of evapotranspiration of water back to the atmosphere depend on the nature of the surface, the availability of water, and the "evaporative demand" of the atmosphere (*i.e.*, the rate at which water vapour can be transported away from the surface under the prevailing meteorologic conditions). Estimation of evapotranspiration rates is important in determining expected rates of stream discharge and in controlling irrigation schemes. The concept of potential evapotranspiration—the possible rate of loss without any limits imposed by the supply of water—has been an important one in the development of hydrology. Most direct measurements of rates of potential evapotranspiration are made using standard evapotranspiration pans with an open water surface. Such measurements serve as a useful standard for comparative purposes, but measured rates may be very different from appropriate potential rates for the surrounding surfaces because of the different thermal and roughness characteristics of the vegetation. In fact, the measured pan rate may be affected by the nature of the surrounding surface due to the influence of evapotranspiration on the humidity of the lower atmosphere.

A distinction also must be drawn between potential rates of evapotranspiration and actual rates. Actual rates may be higher than pan rates for a well-watered, rough vegetation canopy. With a limited water supply available from moisture in the soil, actual rates will fall below potential rates, gradually declining as the moisture supply is depleted. Plants can have some effect on rates of evapotranspiration under dry conditions through physiological controls on their stomata—small openings in the leaf surfaces that are the primary point of transfer of water vapour to the atmosphere. The degree of control varies with plant species.

The only reliable way of measuring actual evapotranspiration is to use large containers (sometimes on the order of several metres across) called lysimeters, evaluate the different components of the water balance precisely, and calculate the evapotranspiration by subtraction. A similar technique is often employed at the catchment scale, although the measurement of the other components of the water balance is then necessarily less precise.

Measuring evapotranspiration

Soil moisture. The soil provides a major reservoir for water within a catchment. Soil moisture levels rise when there is sufficient rainfall to exceed losses to evapotranspiration and drainage to streams and groundwater. They are depleted during the summer when evapotranspiration rates are high. Levels of soil moisture are important for plant and crop growth, soil erosion, and slope stability. The moisture status of the soil is expressed in terms of a volumetric moisture content and the capillary potential of the water held in the soil pores. As the soil becomes wet, the water is held in larger pores, and the capillary potential increases.

Capillary potential may be measured by using a tensiometer consisting of a water-filled porous cup connected to a manometer or pressure transducer. Soil moisture content is often measured gravimetrically by drying a soil sample under controlled conditions, though there are now available moisture meters based on the scattering of neutrons or absorption of gamma rays from a radioactive source.

Methods of measuring soil moisture content

The rate at which water flows through soil is dependent on the gradient of hydraulic potential (the sum of capillary potential and elevation) and the physical properties of the soil expressed in terms of a parameter called hydraulic conductivity, which varies with soil moisture in a nonlinear way. Measured sample values of hydraulic conductivity have been shown to vary rapidly in space, making the use of measured point values for predictive purposes at larger scales subject to some uncertainty.

Water also moves in soil because of differences in temperature and chemical concentrations of solutes in soil water. The latter, which can be expressed as an osmotic potential, is particularly important for the movement of water into plant roots due to high solute concentrations within the root water.

Groundwater. Some rocks allow little or no water to flow through; these are known as impermeable rocks, or aquicludes. Others are permeable and allow considerable storage of water and act as major sources of water supply; these are known as aquifers. Aquifers overlain by an impermeable layer are called confined aquifers; aquifers overlain by an unsaturated, or vadose, zone of permeable materials are called unconfined aquifers. The boundary between the saturated and unsaturated zones is known as the water table. In some confined aquifers, hydraulic potentials may exceed those required to bring the water to the surface. These are artesian aquifers. A well drilled into such an aquifer will cause water to gush to the surface, sometimes with considerable force. Continued use of artesian water, however, will cause potentials to decline until eventually the water may have to be pumped to the surface.

The water found in groundwater bodies is replenished by drainage through the soil, which is often a slow process. This drainage is referred to as groundwater recharge. Rates of groundwater recharge are greatest when rainfall inputs to the soil exceed evapotranspiration losses. When the water table is deep underground, the water of the aquifer may be exceedingly old, possibly resulting from a past climatic regime. A good example is the water of the Nubian sandstone aquifer, which extends through several countries in an area that is now the Sahara desert. The water is being used extensively for water supply and irrigation purposes. Radioisotope dating techniques have

Groundwater recharge

shown that this water is many thousands of years old. The use of such water, which is not being recharged under the current climatic regime, is termed groundwater mining.

In many aquifers, groundwater levels have fallen drastically in recent times. Such depletion increases pumping costs, causes wells and rivers to dry up, and, where a coastal aquifer is in hydraulic contact with seawater, can cause the intrusion of saline water. Attempts have been made to augment recharge by the use of waste waters and the ponding of excess river flows. A scheme to pump winter river flows into the Chalk aquifer that underlies London has reversed the downward trend of the water table.

Measuring water-table levels

Water-table levels in an aquifer are measured by using observation wells. Successive measurements of water levels over time may be plotted as a well hydrograph. The hydraulic characteristics of a particular aquifer around a well can be determined by the response of the water table to controlled pumping. Many aquifers exhibit two types of water storage: primary porosity consisting of the smaller pores and secondary porosity or fractures within the rock mass. The latter may make up only a small proportion of the total pore space but may dominate the flow characteristics of the aquifer. They are of particular importance to the movement of pollutants through the groundwater.

Runoff and stream discharge. Runoff is the downward movement of surface water under gravity in channels ranging from small rills to large rivers. Channel flows of this sort can be perennial, flowing all the time, or they can be ephemeral, flowing intermittently after periods of rainfall or snowmelt. Such surface waters provide the majority of the water utilized by humans. Some rivers, such as the Colorado River in the western United States, are used so intensively that often no water reaches the sea. Others flowing through hot, dry areas, as, for example, the Lower Nile, became smaller downstream as they lose water to evaporation and groundwater storage.

Stream discharge is normally expressed in units of volume per unit time (*e.g.*, cubic metres per second), although this is sometimes converted to an equivalent depth over the upstream catchment area. There are a number of techniques for measuring stream discharge. Measurements of velocities using current meters or ultrasonic sounding can be multiplied by the cross-sectional area of flow. Dilution of a tracer can also be used to estimate the total discharge. Weirs of different types are frequently employed at discharge measurement sites. These are constructed so as to give a unique relationship between upstream water level and stream discharge. Water levels can then be measured continuously, usually with a float recorder, to construct a record of discharge over time—namely, a stream hydrograph. Analysis of the hydrographic response to catchment inputs can reveal much about the nature of the catchment and the hydrologic processes within it.

Techniques for measuring stream discharge

Stream discharge data are presented in terms of daily, monthly, and annual flow volumes, though for some purposes (*e.g.*, flood routing) shorter time periods may be appropriate. The frequency characteristics of peak discharges and low flows are also of importance to water resource planning. These are analyzed using some assumed probability distribution in a way similar to rainfalls. A time recording of annual maximum flood is usually used in flood-frequency analysis. For design purposes the hydrologist may be asked to estimate the flood with a recurrence interval of 50 or 100 years or longer. There are few discharge records that are longer than 50 years, so such estimates are almost always based on inadequate data.

Knowledge of the discharge characteristics of catchments is essential to water supply planning and management, flood forecasting and routing, and floodplain regulation. Discharges vary over short lengths of time during storm periods, seasonally with the seasonal changes in evapotranspiration losses, and over longer periods of time as the rainfall regime changes from year to year. Discharge characteristics also vary with climate. In some places discharge represents only a minor component of the catchment water balance, the losses being dominated by evapotranspiration.

Variations in discharge characteristics

The discharge hydrograph that results from a rainstorm represents the integrated effects of the surface and subsurface flow processes in the catchment. Traditionally, hydrologists have considered the bulk of a storm hydrograph to consist of storm rainfall that has reached the stream primarily by surface routes. Recent work using naturally occurring isotope tracers such as deuterium has shown, however, that in many humid temperate areas the bulk of the storm hydrograph consists of pre-event water. This water has been stored within the catchment between storms and displaced by the rainfall during the storm. This suggests that subsurface flow processes may play a more important role in the storm response of catchments than has previously been thought possible.

Modeling catchment hydrology. The availability of high-speed computers has resulted in a widespread use of computer models in the analysis and prediction of hydrologic variables for research as well as for practical design and management purposes. These models vary greatly in type and complexity, from simple computer implementations of methods previously based on manual calculations to attempts to solve the nonlinear partial differential equations describing surface and subsurface flow processes that require much computation. All have their practical limitations.

Use of computer models for analysis and forecasting

The simpler models treat the catchment as a single "lumped" (or undifferentiated) unit. It is clearly not possible to describe hydrologic processes in detail in such a model, and most processes are represented as empirical functional relationships between inputs and outputs. Some "lumped" models do not refer to the internal hydrologic processes of the catchment at all but use systems analysis techniques to relate inputs to outputs. The parameters of such computer models are calibrated by fitting the model to simulate a known discharge record. It is consequently very difficult to interpret parameters derived in this manner in a physically meaningful way or to extend the use of the model to sites where there are no discharge records. Parameter values for ungauged sites can sometimes be estimated from empirical relationships between catchment characteristics and parameter values derived from fitting a model at a number of gauged sites. The uncertainties in such a procedure, however, are high.

The more complex computer models attempt to analyze the internal processes of the hydrologic system in greater detail, taking into account the spatial nature of the catchment, its topography, soils, vegetation, and geology. These are "distributed" models, usually formulated in terms of flow equations for each hydrologic process considered to be important. Some processes such as channel flows and groundwater flows can be described in a reasonably satisfactory way. In the case of other processes, as, for example, flow through the soil and evapotranspiration, hydrologists cannot be so sure of their descriptions. Distributed models tend to have many parameters. In principle, many of these parameters can be measured in the field or can be estimated from the physical characteristics of the catchment. In practice, such models have proved difficult to apply and have not been shown to provide more accurate results than simpler models in spite of their theoretical rigour.

Most models for hydrologic forecasting in practical use today are deterministic; that is to say, given a sequence of inputs to the model, the outputs are uniquely determined. In a probabilistic description of catchment hydrology, the effects of uncertainty in the model inputs, parameters, or descriptive equations must be reflected in a degree of uncertainty in the outputs. Such a model is known as a stochastic model.

Water quality. Natural water quality is a dilute solution of elements dissolved from the Earth's crust or washed from the atmosphere. Its ionic concentration varies from less than 100 milligrams per litre in snow, rain, hail, and some mountain lakes and streams to as high as 400,000 milligrams per litre in the saline lakes of internal drainage systems or old groundwaters associated with marine sediments.

Factors that affect water quality

Water quality is influenced by natural factors and by human activities, both of which are the subject of much hydrologic study. The natural quality of water varies from place to place with climate and geology, with stream discharge, and with the season of the year. After precipitation

reaches the ground, water percolates through organic material such as roots and leaf litter, dissolves minerals from the soil and rock through which it flows, and reacts with living things from microscopic organisms to humans. Water quality also is modified by temperature, soil bacteria, evaporation, and other environmental factors.

Causes of water pollution

Pollution is the degradation of water quality by human activities. Pollution of surface and subsurface waters arises from many causes, but it is having increasingly serious effects on hydrologic systems. In some areas the precipitation inputs to the system are already highly polluted, primarily by acids resulting from the combustion of fossil fuels in power generation and automobiles.

Other serious causes of pollution have been the dumping of industrial wastes and the discharge of untreated sewage into watercourses. Salt spread on roads in winter has resulted in the contamination of subsurface drinking water supplies in certain areas, as, for example, in Long Island, New York. Excess water resulting from deforestation or irrigation return flows that leach salts from soils in semi-arid areas are major sources of pollution in the western United States and Western Australia.

STUDY OF LAKES

Limnology as a discipline

Limnology is concerned with both natural and man-made lakes, their physical characteristics, ecology, chemical characteristics, internal energy fluxes, and exchanges with the environment. It often includes the ecology and biogeochemistry of flowing freshwaters. The study of former lakes is known as paleolimnology. It involves inferring the history of a former lake basin on the basis of the evidence contained in the sediments of the lake bed.

Lakes may be formed as a result of tectonic activity, glacial activity, volcanism, and by solution of the underlying rock. Man-made lakes or reservoirs may result from the building of a dam within a natural catchment area or as a complete artificial impoundment. In the former case the reservoir may be filled by natural flow from upstream; in the latter the supply of water must be piped or pumped from a surface or subsurface source. The use of reservoir water for water supply, river regulation, or hydroelectric power generation may cause rapid changes in water levels that would not normally occur in a natural lake. In addition, water is usually drawn from a reservoir at some depth, resulting in a shorter residence time relative to an equivalent natural lake.

The history of lakes. A newly formed lake generally contains few nutrients and can sustain only a small amount of biomass. It is described as oligotrophic. Natural processes will supply nutrients to a lake in solution in river water and rainwater, in the fallout of dust from the atmosphere, and in association with the sediments washed into the lake. The lake will gradually become eutrophic, with relatively poor water quality and high biological production. Infilling by sediments means that the lake will gradually become shallower and eventually disappear. Natural rates of eutrophication are normally relatively slow. Human activities, however, can greatly accelerate the process by the addition of excessive nutrients in wastewater and the residues of agricultural fertilizers. The result may be excessive biomass production, as evidenced by phytoplankton "blooms" and rapid growth of macrophytes such as *Eichhornia*.

The physical characteristics of lakes. The most important physical characteristic of the majority of lakes is their pattern of temperatures, in particular the changes of temperature with depth. The vertical profile of temperature may be measured using arrays of temperature probes deployed either from a boat or from a stationary platform. Remote-sensing techniques are being used increasingly to observe patterns of temperature in space and, in particular, to identify the thermal plumes associated with thermal pollution.

Measuring the vertical profile of lake temperature

In summer the water of many lakes becomes stratified into a warmer upper layer, called the epilimnion, and a cooler lower layer, called the hypolimnion. The stratification plays a major role in the movement of nutrients and dissolved oxygen and has an important control effect on lake ecology. Between the layers there usually exists a zone of very rapid temperature change known as the thermocline. When the lake begins to cool at the end of summer, the cooler surface water tends to sink because it has greater density. Eventually this results in an overturn of the stratification and a mixing of the layers. Temperature change with depth is generally much smaller in winter. Some lakes, called dimictic lakes, can also exhibit a spring overturn following the melting of ice cover, since water has a maximum density at 4° C.

Availability of light with depth

A second important characteristic of lakes is the way that the availability of light changes with depth. Light decreases exponentially (as described by Beer's law) depending on the turbidity of the water. At the compensation depth the light available for photosynthetic production is just matched by the energy lost in respiration. Above this depth is the euphotic zone, but below it in the aphotic zone phytoplankton—the lowest level in the ecological system of a lake—cannot survive unless the organisms are capable of vertical migration.

Patterns of sediment deposition in lakes depend on the rates of supply in inflowing waters and on subsurface currents and topography. Repetitive sounding of the lake bed may be used to investigate patterns of sedimentation. Remote sensing of the turbidity of the surface waters also has been used to infer rates of sedimentation, as in the artificial Lake Nasser in Egypt. In some parts of the world where erosion rates are high, the operational life of reservoirs may be reduced dramatically by infilling with sediment.

Water and energy fluxes in lakes. The water balance of a lake may be evaluated by considering an extended form of the catchment water balance equation outlined above with additional terms for any natural or artificial inflows. An energy balance equation may be defined in a similar way, including terms for the exchange of long-wave and shortwave radiation with the Sun and atmosphere and for the transport of sensible and latent heat associated with convection and evaporation. Heat also is gained and lost with any inflows and discharges from the lake. The energy balance equation controls the thermal regime of the lake and consequently has an important effect on the ecology of the lake.

An important role in controlling the distribution of temperature in a lake is played by currents due either to the action of the wind blowing across the surface of the lake or to the effect of the inflows and outflows, especially where, for example, a lake receives the cooling water from a power-generation plant. In large lakes the effect of the Earth's rotation has an important effect on the flow of water within the lake. The action of the wind can also result in the formation of waves and, when surface water is blown toward a shore, in an accumulation of water that causes a rise in water level called wind setup. In Lake Erie in North America, increases in water level of more than one metre have been observed following severe storms. After a storm the water raised in this way causes a seiche (an oscillatory wave of long period) to travel across the lake and back. Seiches are distinctive features of such long, narrow lakes as Lake Zürich, when the wind blows along the axis of the lake. Internal seiche waves can occur in stratified lakes with layers of different density.

Importance of the dissolved oxygen content of lake water

The water quality of lakes. The biological health of a lake is crucially dependent on its chemical characteristics. Limnologists and hydrobiologists are attentive to the dissolved oxygen content of the water because it is a primary indicator of water quality. Well-oxygenated water is considered to be of good quality. Low dissolved oxygen content results in anaerobic fermentation, which releases such gases as toxic hydrogen sulfide into the water, with a drastic effect on biological processes.

Another major concern of limnologists and hydrobiologists is the cycling of basic nutrients within a lake system, particularly carbon, nitrogen, phosphorus, and sulfate. An excess of the latter in runoff waters entering a lake may result in high concentrations of hydrogen ions in the water. Such acid (low values of pH) waters are harmful to the lake biology. In particular, aluminium compounds are soluble in water at low pH and may cause fish to die because of the response induced in their gills.

STUDY OF THE OCEANS AND SEAS

Oceanography and its component disciplines

Oceanography is concerned with all aspects of the Earth's oceans and seas. Physical oceanography is the study of the properties of seawater, including the formation of sea ice, the movement of seawater (*e.g.*, waves, currents, and tides), and the interactions between the so-called World Ocean and the atmosphere. Chemical oceanography is the study of the composition of seawater and of the physical, biological, and chemical processes that govern changes in composition in time and space. Marine geology deals with the geologic evolution of the ocean basins, while biological oceanography or marine ecology focuses on the plant and animal life of the sea.

The origin of the ocean basins. About 71 percent of the Earth's surface is covered by seawater, with the proportion of sea to land being greater in the Southern Hemisphere (four to one) than in the Northern Hemisphere (1.5 to one). Current theories of plate tectonics explain the development of the present ocean basins in terms of a splitting of a large continental landmass brought about by convective circulation within the Earth's mantle. This circulation causes material to rise from the mantle, resulting in the formation of new lithosphere at the mid-oceanic ridges. Continued ascent of material in these areas has resulted in the movement of older material away from the ridges, a process known as seafloor spreading (see also above). Clear evidence of the broad symmetry of these movements has been produced by studies of the residual magnetism of the rocks of the seafloor, which exhibit successive zones of magnetic reversals parallel to the mid-oceanic ridges. Near the continental masses some of the oceanic material sinks into the mantle in zones of subduction, which are associated with oceanic troughs, deep-focus earthquakes, and volcanic activity.

The physical properties of seawater. The physical properties of seawater depend on the chemical constituents dissolved in it. The spatial variability of seawater composition is only partially known, since many areas of the oceans have not been fully sampled. It has been shown that while the salinity of seawater varies from place to place, the relative proportions of the major constituents remain fairly constant. Chlorine accounts for about 55 percent of dissolved solids, sodium 30.6 percent, sulfate 7.7 percent, magnesium 3.7 percent, and potassium 1.1 percent. The average salinity of seawater is about 35 grams of dissolved salts per kilogram of seawater. Higher values occur in areas where evaporation rates are high, such as the Red Sea (41 grams per kilogram), and in areas where relatively pure water freezes out as sea ice, thereby increasing the salinity of the water below. Variations in salinity may be measured indirectly by measurements of the electrical conductivity of seawater, which also yield an accurate estimate of density. With the electrical conductivity method, it is easy to obtain in situ estimates of density and salinity. Using temperature sensors it is also possible to obtain in situ measurements of seawater temperature, the other very important physical characteristic of seawater. It is primarily variations in temperature and density that drive the circulation of water in the oceans.

Measuring the salinity, density, and temperature of seawater

The density of seawater in high latitudes tends to gradually increase with depth. In tropical areas there is usually a well-mixed layer of uniform density close to the surface above the pycnocline, a layer in which density increases extremely rapidly. Below the pycnocline density continues to increase but much more slowly. The pycnocline is a very stable layer that acts as a barrier to the transfer of water and energy between the surface and subsurface layers.

Seawater has a higher specific heat (about 0.95) than land (0.5 on average), so that changes in the temperature of the oceans are much slower than on land. In fact, the oceans represent a vast store of heat that exercises an important, but not fully understood, control on the variability of climate regimes observed on land. The temperature distribution at the surface of the open seas tends to follow lines of latitude, with warmer waters in tropical regions. Close to the landmasses, the lines of equal temperature (isotherms) become deflected, indicating warm and cold coastal currents. Below the surface there is generally a well-mixed layer of fairly uniform temperature about 50 to 200 metres thick below which lies the thermocline, which may extend to depths of 1,000 metres. Below the thermocline, temperature decreases slowly with depth. A typical temperature profile in equatorial regions would be 20° C down to 200 metres, 8° C at 500 metres, 5° C at 1,000 metres, and 2° C at 4,000 metres.

Plotting the temperature and salinity of a sample of seawater on a graph with linear axes (a T–S diagram) can be a powerful research tool. A mass of fully mixed water having a homogeneous distribution of temperature and salinity would plot as a single point on a T–S diagram. For actual water masses it is common to find that points plotted for samples taken from different depths plot as a curve (sometimes complex) on the diagram. Such curves provide an indication of the mixing between different water masses that is taking place in the profile. T–S curves also are useful for uncovering errors in data.

T–S diagrams

The circulation of the oceans. One major cause of the circulation of waters in the oceans is the difference in the energy budget between the tropics and the poles (the thermohaline circulation). While it is now thought that differences in solar heating have a relatively minor direct effect on ocean circulation, the formation of sea ice and the loss of heat from the oceans at the poles causes a movement of colder, denser water toward the Equator at depth. The major surface currents of the oceans are driven by the surface shear stresses imposed by the wind. These motions are influenced by the topography of the ocean basins and the Coriolis effect due to the Earth's rotation, so that in the Northern Hemisphere the moving water becomes deflected toward the right, while in the Southern Hemisphere toward the left. This results in major clockwise circulations in the North Atlantic (including the Gulf Stream) and the North Pacific, with counterclockwise circulations in the South Atlantic, South Pacific, and Indian oceans. Within the tropics there tends to be a pattern of westward-flowing currents in both the Northern and Southern hemispheres, with an eastward-flowing countercurrent close to the Equator itself. There may also be an eastward-flowing undercurrent at depth. It is certain that there is still much to be learned about the details of ocean circulation, particularly at depth.

There are two fundamental approaches to measuring ocean currents: the Lagrangian and Eulerian methods. In the Lagrangian method individual parcels of water are tracked using floats or buoys. Satellite-tracked buoys equipped with radio transmitters are now commonly used in the study of surface currents. Currents at depth may be studied with Swallow floats, which are adjusted to be neutrally buoyant at a certain density of seawater. Tracer techniques, such as those involving the use of dyes and discharges of pollutants, may also be employed to track flowing water at least in coastal areas. The greatest number of measurements of surface currents by the Lagrangian method, however, have come from the records of the drift of ships contained in navigation logs.

The Lagrangian and Eulerian methods of measuring ocean currents

The Eulerian method consists of measuring the velocity of flow past a fixed point (a moored ship, anchored line, or structure) with a current meter, of which there are a number of different types. Flow velocities may be measured as a function of both depth and time at any site.

An indirect method for estimating current velocities is the geostrophic method. It is based on the fact that the movement of water masses away from the sea surface and any solid boundary can be assumed to be frictionless and unaccelerated. Under such conditions the pressure gradient and the effects of gravity and Coriolis forces should balance exactly. The expected rates and directions of flow can then be computed theoretically. It has been shown that the geostrophic currents are good approximations of actual currents.

Geostrophic method

The hydrodynamics of ocean currents can be described by the dynamic equations of fluid flow. The advent of high-speed digital computers has made it possible to obtain approximate numerical solutions to these equations for many problems of practical interest, including the transient effects of tides. The formation and propagation of waves, together with their refraction in shallow coastal waters, also can be computed numerically.

Biogeochemical cycles in the oceans. The ocean is a great store of chemicals that receives inputs from rivers and the atmosphere and, on average, loses equal amounts to sedimentary deposits on the ocean floor. Biological processes play a large part in processing the chemicals received and in maintaining the remarkable consistency in the composition of seawater. Fortunately this consistency does not extend to all the elements found in seawater. Concentrations of some of the minor, or trace, elements can be used to infer the mixing, biological, and sedimentation processes that occur. Throughout the oceans the major variations in composition are in the upper layers, where the greatest biological activity is found.

Radiometric dating

The use of a number of different radioisotopes in dating sediments and calculating rates of sedimentation and mixing within the oceans have been important in studying the biogeochemical cycles of the oceans. A particularly interesting use of radiometric dating was in investigating the formation of the manganese nodules that occur on certain segments of the seabed and in the underlying sediments. These nodules consist primarily of manganese and iron oxides, even though concentrations of these elements in seawater are very low. Dating techniques have shown that the growth rates of the nodules are on the order of three millimetres per 1,000 years, or 1,000 times less than the accumulation rate of the sediments on which they lie.

Remote sensing of the oceans. One of the fundamental problems faced by oceanographers is the sheer size of the oceans and the consequent need to rely on special surface vessels and submersibles for direct measurements. It can be very costly to operate either type of vessel on long deep-sea expeditions. Moreover, observations from such craft can provide only a partial picture of oceanic phenomena and processes in terms of both space and time. Consequently, there has been considerable interest in taking advantage of remote-sensing techniques in oceanography, particularly those that use satellites. Remote sensing allows measurements to be made of vast areas of ocean repeated at intervals in time.

Satellite observations

The first satellite devoted to oceanographic observations was Seasat, which orbited the Earth for three months in 1978. Its polar orbit made it possible to provide coverage of 95 percent of the major oceans every 36 hours. Seasat carried radiometers for observations at visible, infrared, and microwave wavelengths, along with radar scatterometers, imaging radar, and an altimeter. This array of instruments yielded much data, including an estimation of sea-surface temperatures, net radiation inputs to the sea surface, wave heights, and wind speeds close to the sea surface. In addition, patterns of near-surface sediment movement and other information were derived from an analysis of the satellite images. For further information about remote-sensing techniques used in oceanographic research, see EXPLORATION: *Acoustic and satellite sensing.*

STUDY OF ICE ON THE EARTH'S LAND SURFACE

The scope of glaciology

Glaciology deals with the physical and chemical characteristics of ice on the landmasses; the formation and distribution of glaciers and ice caps; the dynamics of the movement of glacier ice; and interactions of ice accumulation with climate, both in the present and in the past. Glacier ice covers only about 10 percent of the Earth's land surface at the present time, but it was up to three times as extensive during the Pleistocene Ice Age.

The accumulation of ice. Glacier ice forms from the accumulation of snow over long periods of time in areas where the annual snowfall is greater than the rate of melting during summer. This accumulated snow gradually turns into crystalline ice as it becomes buried under further snowfalls. The process can be accelerated by successive melting and freezing cycles. The crystalline ice incorporates some of the air of the original snow as bubbles, which only disappear at depths exceeding about 1,000 metres. Successive annual layers in the ice often can be distinguished by differences in crystalline form, by layers of accumulated dust particles that mark each summer melt season, or by seasonal differences in chemical characteristics such as oxygen isotope ratios. The layers become thinner with depth as the density of the ice increases.

Measurements of oxygen isotope ratios

Oxygen isotope ratios indicate the temperature at which the snow making up the ice was formed. Seasonal variations in isotope ratios not only allow annual layers to be distinguished but also can be used to determine the residence times of melt waters within an ice mass. Long-term variations in isotope ratios can be employed to ascertain temperature variations related to climatic change. An ice core of 1,390 metres taken at Camp Century in Greenland has been used in this way to indicate temperatures during the past 120,000 years, and it shows clearly that the last glacial period extended from 65,000 to about 10,000 years ago. These results have been corroborated by measurements of additional cores from Greenland and Antarctica. In spite of the fact that temperatures may remain below freezing throughout the year, ice accumulation over much of Antarctica is very slow, since precipitation rates are low (they are equivalent to those in many desert areas).

On any glacier there is a long-term equilibrium between accumulation and ablation (losses due to melt runoff and other processes). Continued accumulation eventually causes ice to move downhill, where melt rates are higher. The elevation at which accumulation balances losses changes seasonally as well as over longer periods. In many areas of the world, the annual meltwaters are a crucial part of the water resources utilized by man. In the past it was very difficult to predict amounts of spring melt runoff because of the difficulties in assessing snow accumulation in mountainous terrain. Remote-sensing techniques now allow accumulation over much larger areas to be estimated, and they also offer the possibility of updating those estimates during the melt season.

The movement of glaciers. The mechanisms by which a large mass of ice can move under the effects of gravity have been debated since about 1750. It is now known that some of this movement is due to basal sliding but that the ice itself, a crystalline solid close to its melting point, can flow, behaving like other crystalline solids such as metals. Early measurements of flow velocities were based entirely on surveys of surface stakes, a technique still used today. During the early 19th century the Swiss geologist Louis Agassiz showed that the movement was fastest in the central part of a glacier. Rates of movement are fastest in the temperate glaciers, which have temperatures close to the melting point of ice and include about 1 percent liquid water. (This water constitutes a layer at the bottom of such an ice mass.) Velocities vary through time, quite dramatically at times. Certain glaciers (*e.g.*, the Muldrow and Variagated glaciers in Alaska) are subject to surges of very rapid velocities at irregular periods. The causes of these catastrophic advances are still not well understood.

Modern techniques of studying the movement of glacial ice

Techniques for investigating the movement of ice in the field include studies of the deformation of vertical boreholes and lateral tunnels dug into the ice. The internal structure of glaciers and the Greenland and Antarctic ice caps have also been examined by means of radar sounding. This method works best in cold glaciers where the ice is below its freezing point.

Indirect evidence of the patterns of movement is obtained from the characteristic landforms associated with glaciers, particularly scratched or striated bedrock and moraines composed of rock debris. Such forms also allow the interpretation of former patterns of movement in areas no longer covered by ice.

PRACTICAL APPLICATIONS

Development and management of water resources. Water is essential to many of humankind's most basic activities—agriculture, forestry, industry, power generation, and recreation. As the hydrologic sciences provide much of the knowledge and understanding on which the development and management of available water resources are based, they are of fundamental importance.

In 1965 the United Nations Educational, Scientific and Cultural Organization (UNESCO) initiated the International Hydrological Decade (IHD), a 10-year program that provided an important impetus to international collaboration in hydrology. Considerable progress was made in hydrology during the IHD, but much still remains to be done, both in the basic understanding of hydrologic pro-

cesses and in the development and conservation of available water resources. Many developing countries remain highly susceptible to diseases related to a lack of water supplies of good quality and to the effects of drought. This has been cruelly highlighted in recent times by the severe droughts in the Sahel region of Africa in the periods 1969–74 and 1982–85 (see below).

In the developed countries the ready availability of a supply of good quality water is expected. Yet, even in the most advanced countries, many water sources are not being used wisely. Groundwater levels in certain areas have fallen dramatically since the 1940s, leading to ever higher pumping costs. Other surface and subsurface water sources are becoming increasingly polluted by urban, agricultural, and industrial wastes in spite of increased expenditure on waste-water treatment and legislation of minimum quality standards. Humankind continues to use the oceans as a vast dumping ground for its waste products, even though little is known about the effects of such wastes on marine ecosystems. It is no exaggeration to say that the misuse of water resources will become a major source of conflict between communities, states, and nations in the years to come. Already several disputes over rights to clean water have taken on international significance.

Misuse of water resources

Since the early 1980s the acid rain problem has assumed scientific, economic, and political prominence in North America and Europe. This major environmental problem serves to illustrate the interdependence of the various hydrologic sciences with other scientific disciplines and human activities. As was noted earlier, waste gases (primarily oxides of sulfur and nitrogen) enter the atmosphere from the burning of fossil fuels by automobiles and electric power plants. These gases combine with water vapour in the atmosphere to form sulfuric and nitric acids. When rain or some other form of precipitation falls to Earth, it is highly acidic (often with a pH value of less than 4). The resultant acidification of surface and subsurface waters has been shown to have detrimental effects on the ecology of affected catchments. Areas underlain by slowly weathering bedrock, such as in Scandinavia, the Adirondack Mountains of New York, and the Canadian Shield in Quebec are particularly susceptible. Many lakes in these areas have been shown to be biologically dead. There also is evidence that the growth of trees may be affected, with consequent economic ramifications where forestry is a major activity. The areas most greatly affected may be far downwind of the source of the pollution. A number of countries have claimed that the major sources of acid rain affecting their streams and lakes lie outside their borders.

Acidification of surface and subsurface waters

Research has revealed that in an area susceptible to the effects of acid rain short-lived events can have a particularly damaging effect. These "acid shocks" may be due to inputs of highly acid water from a single storm or to the first snowmelt outflows in which the major part of the pollutant input accumulated over the winter is concentrated. The way in which the chemistry of the input water is modified in its flow through the catchment depends both on the nature of the soils and rocks in the catchment and on the flow paths taken through the system. These interactions are at present poorly understood. It is likely, however, that the attempt to understand the chemical processes within the different flow paths will lead to significant improvements in scientific understanding of catchment hydrology.

Concern over groundwater quantity and quality. Groundwater problems are becoming increasingly serious in many areas of the world. Rapid increases in the rates of pumping of groundwater in many aquifers has caused a steady lowering of water table levels where extraction has exceeded rates of recharge. A notable example is the Ogallala aquifer, a sandy formation some 100 metres thick, which lies beneath the Great Plains from South Dakota to Texas. It has been estimated that as much as 60 percent of the total storage of this huge aquifer has already been extracted primarily for agricultural use. The remaining water, if it continues to be mined in this way, will become more and more expensive to extract. This situation points out the importance of understanding groundwater flow and recharge processes in complex heterogeneous formations so that safe yields of aquifers can be properly predicted.

There are many causes of groundwater pollution; most are the accidental or incidental consequences of human activities (*e.g.*, pollution resulting from the use of artificial fertilizers or saltwater intrusion into coastal aquifers due to excessive pumping). In some cases, however, groundwater may be contaminated because of planned human effort. Subsurface repositories of water, for example, have been considered as possible receptacles for waste products, including radioactive materials. This has resulted in both experimental and model studies of water flows in poorly permeable massive rocks that would be used to store such wastes. The effects of joints and fractures on the very slow transport of contaminants over long periods of time in such rocks is as yet uncertain but must be clarified if this form of storage is to be proved safe.

Studying the causes of droughts and other climatic patterns. Another subject still poorly understood is the occurrence of droughts in areas of highly variable rainfall. In the early 1970s and again in the early 1980s the Sahel region of Africa suffered periods of severe drought, resulting in widespread famine and death. There have been many Sahelian droughts before, but the consequences of the recent droughts have been exacerbated by increased populations of people and grazing animals. The combination of drought and population growth results in desertification. It remains an unanswered scientific question as to whether the deterioration of the Sahel and other marginal lands is part of a long-term natural change or whether it is a result of human activities.

Some evidence for long-range interactions in the occurrence of droughts and other climatic regimes comes from studies of the ocean currents. It is known that the oceans are a major controlling influence on climate, but the processes involved remain the subject of active research. Some clues have been revealed by studies of El Niño, a minor branch of the Pacific Equatorial Countercurrent that flows south along the coasts of Colombia and Ecuador where it meets the northward-flowing Peru Current. The cold Peru Current keeps rainfall along the coastal area of Peru very low but maintains a rich marine life, which in turn supports major bird populations and a fishing industry. In certain years El Niño becomes much stronger, forcing the Peru Current to the south. Storms rake the coast, causing flooding and erosion. The sudden change in sea temperatures causes dramatic decreases in plankton production and, consequently, in fish and bird populations. Catastrophic El Niño events occurred in 1925, 1933, 1939, 1944, 1958, and 1983. It is thought that the global changes associated with this last event included severe droughts in Australia and Central America and floods in the southwestern United States and Ecuador. Explanations of the El Niño events have invoked both local and long-range interactions in the circulation of the Pacific winds and currents. The study of such dramatic events, enhanced by remote sensing and computer modeling, is a major stimulus to understanding the general circulation of the Earth's atmosphere and oceans. (K.J.B.)

Studies of the effects of El Niño on global climate

Atmospheric sciences

The atmospheric sciences focus on the structure and dynamics of the Earth's atmosphere. Such mathematical tools as differential equations and vector analysis, together with large computer systems, are used to evaluate the physical and chemical relations that describe the workings of the atmosphere. Planetary science and stellar science are comparable fields of study of the atmospheres of other astronomical bodies.

Scope of the atmospheric sciences

The atmospheric sciences traditionally have been divided into three topical areas—meteorology, climatology, and aeronomy. In meteorology the focus of study is the day-to-day, hour-to-hour changes in weather within the lower stratosphere and troposphere. Climatology, on the other hand, concentrates on a statistical description of the weather in the same region of the atmosphere but over long periods of time ranging from a month to millions of years. Meteorology is the science from which weather

forecasts are prepared, whereas climatology provides an explanation of why climate differs across the Earth and how it is interrelated with other components of the natural environment. Studies of the atmospheric regions above the lower stratosphere are associated with the field of aeronomy. They deal with such matters as the photochemical processes of the upper atmosphere, ionospheric physics, airglow, magnetospheric storms, and auroral phenomena.

The scope of the atmospheric sciences is indeed broad. Here it is possible only to survey the basic concerns and principal areas of research of the various disciplines involved.

STUDY OF THE EVOLUTION OF THE ATMOSPHERE

The evolution of the Earth's atmosphere remains incompletely understood. Investigators believe that the present atmosphere resulted from a gradual release of gases from the Earth's interior and that it is quite distinct from the primordial atmosphere, which developed by outgassing during the formation of the planet. Current volcanic gaseous emissions include carbon dioxide, sulfur dioxide, chlorine, fluorine, water, and diatomic nitrogen as well as traces of other substances. Approximately 85 percent of the emissions is in the form of water vapour. Carbon dioxide constitutes about 10 percent of the effluent.

Researchers hold that water must have been able to exist in its liquid state during the early evolution of the atmosphere since the oceans appear to have been present for at least 3,000,000,000 years. Because the solar output was about 25 percent less 4,000,000,000 years ago, enhanced levels of carbon dioxide and perhaps ammonia apparently were required to retard long-wave radiative heat loss to space. The initial life forms that evolved in this environment must have been anaerobic (*i.e.*, capable of surviving in the absence of oxygen) and had to have been able to resist the biologically destructive short ultraviolet radiation that was not absorbed by a layer of ozone (O_3) as it is today. Once organisms developed the capability for photosynthesis (in which visible sunlight on plants produces diatomic oxygen [O_2]), oxygen was produced in large quantities, eventually reaching the current levels. This also permitted the ozone layer to develop, as O_2 was dissociated into monatomic oxygen and recombined with O_2 to form ozone. The capability of primitive plant forms to photosynthesize developed between 2,000,000,000 and 3,000,000,000 years ago. Prior to the emergence of photosynthetic organisms, oxygen was produced in only limited quantities through the decomposition of water vapour into molecular oxygen by ultraviolet radiation from the Sun.

Present composition of the atmosphere

The present composition of the Earth's atmosphere, in terms of total molecules, is diatomic nitrogen (78.08 percent), diatomic oxygen (20.95 percent), argon (0.93 percent), water (from about 0 to 4 percent), and carbon dioxide (0.0325 percent). The inert gases neon, helium, and krypton and other constituents such as nitrogen oxides and sulfur compounds are found in lesser amounts.

STUDY OF SURFACE BUDGETS

Heating by radiation, conduction, and convection. The Earth's atmosphere is bounded at the bottom by water and land. Heating of this surface is accomplished by three physical processes: radiation, conduction, and convection.

Incoming energy

The electromagnetic radiation that influences both the temperature of the atmosphere and that of the surface is commonly divided into two types: insolation from the Sun, which is referred to as shortwave radiation (with predominant wavelengths of 0.39 to 0.76 micrometre); and emittance from the surface and the atmosphere, which is called long-wave radiation (with typical wavelengths of four to 30 micrometres). The wavelength of the emitted electromagnetic radiation depends on the temperature of the radiating body as specified by Planck's law. The Sun, with its surface temperature of around 6,000 K, emits at a much shorter wavelength than the Earth, which has surface and atmospheric temperatures of about 250 to 300 K.

A fraction of shortwave radiation is absorbed by atmospheric gases and warms the air directly. Most of this energy, however, reaches the surface when clouds are not present. Scattering of a fraction of the shortwave radiation, particularly of the shortest wavelengths by air molecules (known as Rayleigh scattering), produces blue skies.

When clouds are present, a large percentage (up to about 80 percent) of the solar insolation is reflected into space (the fraction of back-reflected, shortwave radiation is called the albedo). Of the solar radiation reaching the surface, a fraction is reflected into the atmosphere. Values of the surface albedo range from 0.95 for fresh snow to 0.10 for dark organic soils. On land this reflection occurs entirely at the surface. In water, however, shortwave radiation penetrates to significant depths (as much as several hundred metres) before the insolation is completely attenuated. The heating by solar radiation in water, therefore, is distributed through depth, which results in smaller temperature changes at a given level than would occur with the same insolation over land at the surface.

The magnitude of solar radiation reaching the surface depends on latitude, time of year, time of day, and orientation of the land surface with respect to the Sun. In the Northern Hemisphere north of 23°30′, for example, solar insolation at local noon is less on north-facing slopes than on land oriented toward the south.

Solar radiation consists of two components: direct and diffuse radiation. Direct shortwave radiation is that which reaches a point without being absorbed or scattered from its line of propagation by the intervening atmosphere. The image of the Sun's disk as a sharp, distinct object represents that portion of the solar radiation that reaches the viewer directly. Diffuse radiation, in contrast, reaches the observer only after being scattered from its line of propagation. On an overcast day, for example, the Sun's disk is not visible and all of the shortwave radiation is diffuse.

Long-wave radiation is emitted by the atmosphere and propagates both upward and downward. The magnitude of this radiation reaching the surface depends on the temperature at the height of emission and the amount of absorption between the height of emission and the surface. A larger fraction of the long-wave radiation is absorbed when the intervening atmosphere has large amounts of water vapour and carbon dioxide. Clouds that have a liquid water content on the order of 2.5 grams per cubic metre absorb almost 100 percent of the long-wave radiation within a depth of 12 metres into the cloud. Clouds with less water content require greater depths before complete absorption is attained (*e.g.*, a cloud with a water content of 0.05 grams per cubic metre requires about 600 metres for complete absorption). Clouds that are at least this thick emit long-wave radiation from their base, corresponding to the temperature of the lowest levels of the cloud.

Heat flux by conduction

The magnitude of heat flux by conduction below the surface depends on thermal conductivity and the vertical gradient of temperature in the material beneath the surface. Such soils as dry peat, which has very low thermal conductivity, permit little heat flux, whereas concrete, which has a thermal conductivity almost 100 times as great, allows substantial heat flux. In water, thermal conductivity is relatively unimportant because, in contrast to land surfaces, solar insolation extends to substantial depths into the water, and water can be mixed vertically.

Vertical mixing (convection) occurs in the atmosphere as well as in water. Also referred to as turbulence, this mechanism of heat flux occurs in the atmosphere in two forms. When the surface is warmer than the overlying air, mixing occurs spontaneously to redistribute the heat. This process, termed free convection, occurs when the atmospheric lapse rate of temperature decreases at a rate greater than −0.98° C per 100 metres, the adiabatic lapse rate. In the ocean the temperature increase with depth, which can result in free convection, is dependent on temperature, salinity, and depth. At the surface with a temperature of 20° C and a salinity of 34.85 parts per 1,000, an increase of temperature with depth of more than about 0.19° C per kilometre results in free convection.

Free and forced convection

Mixing also can occur because of the shearing stress of the wind on the surface. As a result of surface friction, the average wind velocities at the surface must be zero unless the surface is moving. Winds above the surface are decelerated when the vertical shear of the wind becomes large enough to cause vertical mixing. This process by which

heat and other atmospheric properties are mixed as a result of wind shear is called forced convection. (Free and forced convection are also referred to as convective turbulence and mechanical turbulence, respectively.) Forced convection occurs either as sensible turbulent heat flux in which heat is directly transported to or from the surface or as latent turbulent heat flux in which heat is used to evaporate water from the surface.

The temperature at the interface of the atmosphere and the surface results from the contributions of heat by radiation, conduction, and convection. The magnitude of these contributions depends on the wind, temperature, and moisture structure in the immediate overlying atmosphere, on the intensity of solar insolation, and on the physical characteristics of the surface.

Other factors affecting atmospheric processes and conditions. The water budget at the air–surface interface is also of crucial importance in influencing atmospheric processes. The surface gains water through precipitation (rain and snow) and by direct condensation and deposition (dew and frost). On land precipitation is often heavy enough for some of it to percolate into the ground or flow as runoff into streams, rivers, lakes, and the oceans. A portion of the precipitation that remains on the surface, such as in puddles or on vegetation, immediately evaporates back into the atmosphere.

Evapotranspiration and evaporation

Liquid water is also converted to water vapour by evapotranspiration as vegetation extracts water from the soil and emits it through stoma on the leaves and by evaporation directly from the surface of the soil when water from below is diffused upward. Evaporation occurs at the surface of water bodies at a rate inversely proportional to the relative humidity just above the surface. Evaporation is rapid in dry air but much slower when the lowest levels of the atmosphere are almost saturated. Evaporation from soil is dependent on the rate at which moisture is supplied by capillary suction within the soil, while evapotranspiration is dependent on the water available to plants within the root zone and whether or not the stoma are open on the leaf surfaces. Water that evaporates and evapotranspires into the atmosphere is often transported long distances over the Earth before it is precipitated.

The input, transport, and removal of water from the atmosphere is part of the hydrologic cycle (see above). At any one time only a very small fraction of the Earth's water is present within the atmosphere. If it were all condensed out, it would only cover the surface of the Earth to an average of about 2.5 centimetres.

Investigators have determined surface budgets for other constituents of the atmosphere as well. The nitrogen budget, for example, involves the chemical transformation of diatomic nitrogen (N_2), which makes up 78 percent of the atmospheric gases, into compounds containing ammonium (NH_4^+), nitrite (NO_2^-), and nitrate (NO_3^-). This conversion, referred to as nitrification, is performed by *Rhizobium* and other nitrifying bacteria that live on the roots of legumes such as peas and clover. Lightning is also a nitrifying agent. The nitrogen compounds are eventually reconverted to N_2 after the plants die or are eaten by denitrifying bacteria. These bacteria, in their consumption of plants and the excreta and corpses of plant-eating animals, perform much of this reconversion to N_2. Some of the compounds, however, are changed back to N_2 by a series of chemical processes associated with ultraviolet light from the Sun. The combustion of petroleum by motor vehicles also produces oxides of nitrogen, which has enhanced the natural concentrations of these compounds. Part of the smog in urban areas is associated with substantially higher levels of these nitrogen compounds.

The sulfur budget is another atmospheric constituent of major importance. Sulfur enters the atmosphere as a result of the weathering of sulfur-containing rocks and by intermittent volcanic emissions. Organic forms of sulfur are incorporated into living organisms and represent an important component in the structure and function of proteins. Sulfur also appears in the atmosphere in the form of the gas sulfur dioxide (SO_2) and as part of particulate matter containing sulfate (SO_4). These forms of sulfur are dry deposited directly or are precipitated onto the Earth's surface. When they are wet, these sulfur compounds transform into caustic sulfuric acid (H_2SO_4).

During the last century significant quantities of sulfur have been released into the atmosphere through the combustion of fossil fuels. In and around regions of urbanization and heavy industrial activity, the enhanced precipitation and deposition of sulfur as sulfuric acid and of nitrogen oxides as nitric acid have been associated with damage to fish populations, forests, and the exteriors of buildings and statues. This is commonly called the acid rain problem. (For details, see above *Development and management of water resources*.)

Carbon budget

The carbon budget in the atmosphere is of critical importance to climate and to life. Carbon appears in the Earth's atmosphere primarily as carbon dioxide (CO_2), which is produced naturally by the respiration of living organisms, during the decay of these organisms, through the weathering of carbon-containing rock strata, and from volcanic emissions. Plants utilize CO_2, water, and solar insolation to convert CO_2 to diatomic oxygen. This process, called photosynthesis, results in about a 3 percent drop in CO_2 concentrations in the Northern Hemisphere during the growing season (spring to fall). CO_2 is also absorbed by the ocean waters with the rate of exchange to the ocean greater for colder waters. Currently CO_2 constitutes about 0.03 percent of the gaseous composition of the atmosphere. In past geologic times CO_2 levels are thought to have been significantly higher than they are today and to have had a significant effect on climate and ecology. During the Carboniferous Period (from about 360,000,000 to 285,000,000 years ago), for example, moderately warm and humid climates and high concentrations of CO_2 were associated with extensive lush vegetation. After these plants died and decomposed, they were converted to sedimentary rocks and became the coal deposits currently used by industry.

The so-called greenhouse effect

In the atmosphere certain wavelengths of long-wave radiation are absorbed and then reemitted by CO_2. Since the lower levels of the atmosphere are warmer than higher layers, the absorption of upward-propagating electromagnetic radiation and the reemission of a portion of it back downward permit the lower atmosphere to remain warmer than it would be otherwise. Erroneously referred to as the greenhouse effect (a greenhouse retains heat primarily because solar radiation enters through the glass, but mixing of air into the greenhouse from above is constrained by the glass), higher concentrations of CO_2 in the air appear to be associated with a warmer lower troposphere.

STUDY OF THE VERTICAL STRUCTURE OF THE ATMOSPHERE

The atmosphere is divided into several distinct layers, which are defined on the basis of whether the temperature increases or decreases with height. The lowest major layer is the troposphere, in which the temperature generally decreases with height. This layer contains most of the Earth's clouds and is the region in which what is known as weather primarily occurs.

The planetary boundary layer. *Characteristics.* The lower levels of the troposphere are usually strongly influenced by the Earth's surface. Known as the planetary boundary layer, they compose the region of the atmosphere in which the surface influences temperature, moisture, and velocity through the turbulent transfer of mass. As a result of surface friction, winds in this layer are usually weaker than they are above that height and blow toward areas of low pressure. The study of the atmospheric structure and dynamics within the layer is referred to as boundary layer meteorology, or micrometeorology.

Micrometeorology

Under clear, sunny skies over land the planetary boundary layer tends to be relatively deep because of the heating of the ground by the Sun and the resultant generation of convective turbulence. During the summer the boundary layer can reach, for example, heights of one to 1.5 kilometres in the eastern United States and up to five kilometres in the southwestern desert region of the country. Under this condition the temperature decreases at the dry adiabatic lapse rate (−9.8° C per kilometre) throughout most of the boundary layer, with a superadiabatic lapse rate of a greater magnitude near the heated surface. In contrast, during clear, calm nights, turbulence tends to cease and

radiational cooling from the surface results in a temperature that increases with height above the surface.

The gradient Richardson number, defined as

$$Ri = \frac{g}{T}\left[\frac{\Delta T}{\Delta z} + 9.8^\circ \text{ C/km}\right] / \left(\frac{\Delta \vec{V}}{\Delta z}\right)^2,$$

is a commonly used parameter to describe the tendency of the atmosphere to be turbulent. In this expression $\Delta T/\Delta z$ is the temperature lapse rate, $\Delta \vec{V}/\Delta z$ is the velocity shear in the vertical, and g is the gravitation acceleration. The parameter Ri represents the ratio of the generation or suppression of turbulence by buoyant production of energy to the mechanical generation of energy by wind shear.

When a region of the atmosphere has a temperature decrease with height, $\Delta T/\Delta z$, greater in magnitude than the adiabatic lapse rate, turbulence is generated by convective overturning as the warmer, lower level air rises and mixes with the cooler air aloft. Since the environmental lapse rate is greater in magnitude than the adiabatic lapse rate, an ascending parcel of air remains warmer than the surrounding ambient air even though the parcel undergoes expansional cooling. This overturning occurs in the form of bubbles (eddies) of warmer air. The larger bubbles often have sufficient buoyant energy to penetrate the top of the boundary layer and entrain air from aloft into the layer, thereby deepening it. Since, in general, the air aloft has a lapse rate that is less in magnitude than the adiabatic lapse rate, compressional warming of this entrained air results in a heating of the boundary layer. The top of the daytime boundary layer is called the mixed-layer inversion.

The ability of the convective bubbles to penetrate the top of the boundary layer depends on the temperature lapse rate aloft. Since the numerator of Ri is usually positive at those heights, the turbulence of penetrative convective bubbles is rapidly eliminated as the parcel quickly becomes cooler than the ambient environment and negatively buoyant with additional ascent. The height that the boundary layer attains on a sunny day, therefore, is strongly influenced by the intensity of surface heating (which, with strong heating, results in large negative values of Ri near the surface) and the temperature lapse rate just above the boundary layer. The less negative the value of $\Delta T/\Delta z$ above the boundary layer, the greater is the suppression of turbulent bubble penetration.

On clear, calm nights the value of Ri becomes large and positive as radiational cooling results in a temperature increase with height. Turbulence is suppressed by the strong thermal stratification, and nearly laminar flow can result over flat terrain. The depth of the radiationally cooled layer, or nocturnal inversion, depends on factors that include the moisture content of the air, soil and vegetation characteristics, and terrain configuration. In a dry desert, for instance, the nocturnal inversion is higher than in a moister environment. The inversion in a humid environment is lower because more upward-propagating longwave radiation is absorbed by the greater number of water molecules and reemitted downward, thereby preventing the lower levels from cooling as rapidly.

Nocturnal inversion

During windy conditions the mechanical production of turbulence, expressed by the denominator in Ri, becomes important. Turbulence eddies produced by wind shear tend to be smaller in size than the turbulence bubbles produced by buoyancy. Within a few tens of metres from the surface during windy conditions, the wind speed is very accurately represented as a logarithmic function of height. If the winds are sufficiently strong, even with a positive value of Ri, the generation of turbulence by wind shear can dominate the dissipation of turbulence by the stable temperature stratification. Theoretically, the wind shear has to be such that $Ri < 0.25$ for this mechanical turbulence to exist, otherwise laminar flow will result.

Above the boundary layer in the troposphere, Ri tends to be greater than 0.25. The exceptions are near jet streams where large velocity shears exist, in and adjacent to cumuliform clouds where buoyant turbulence is generated as a result of the release of latent heat, and at and just above cloud tops where radiational cooling from the clouds causes a destabilization and generates buoyancy.

Monitoring the planetary boundary layer. Studies of the boundary layer involve both in situ measurements and remote-sensing observations. Meteorologic measurements of the first type include those of temperature, pressure, wind speed, shortwave and long-wave radiative fluxes, and the composition of air. They are conducted with the aid of towers, tethered balloons, and surface data-collection platforms. An assortment of remote-sensing observations are made with active systems, such as Doppler and non-Doppler radars, lidars (light radars), and acoustic sounders. The radar systems measure the back scattering of microwave radiation with wavelengths on the order of three to 10 centimetres. Non-Doppler radar provides estimates only of precipitation intensity, while the Doppler variety can furnish estimates of wind speed and direction as well. Shorter wavelength Doppler radar is often able to measure winds even in clear air. The carbon dioxide infrared lidar—a pulsed laser radar of a wavelength of 10.6 micrometres—can measure wind structure and turbulence within a few tens of kilometres of the instrument. Acoustic sounders enable investigators to monitor the depth and structure of the boundary layer using echo return characteristics. Some remote-sensing observations are made with passive systems, including a pyranometer, which measures direct and diffuse radiation from the Sun, and a pyrheliometer, which samples direct solar radiation only.

Remote-sensing systems

The free atmosphere and the tropopause. *General characteristics and cloud forms.* The region above the boundary layer is commonly referred to as the free atmosphere. Here, winds are not directly retarded by surface friction. Clouds occur most frequently in this portion of the troposphere (the exceptions are fog and clouds that impinge on or develop over elevated terrain).

There are two basic types of clouds: stratiform and cumuliform. Both cloud types develop when clear air ascends, cooling adiabatically as it expands until water begins to condense or deposition occurs. This change in the state of water occurs because cooler air can hold less water than can warmer air.

Classification of clouds

Stratiform clouds form as saturated air is mechanically forced upward, remaining colder than ambient clear air at the same height. In the lower troposphere such clouds are called stratus. Advection fog is a form of stratus cloud whose base lies at the Earth's surface. In the middle troposphere stratiform clouds are called altostratus, while in the upper troposphere the terms cirrostratus and cirrus are applied. The latter refers to thin, often wispy cirrostratus. Precipitating stratiform clouds that extend through a large part of the troposphere are known as nimbostratus.

Cumuliform clouds occur when saturated air is turbulent. Such clouds, with their bubbly, turreted shapes, make it possible to view small-scale, up-and-down motions similar to those that occur, but which are not visually observable, in an unsaturated turbulent planetary boundary layer. Cumulus clouds are often seen with bases near or at the top of the boundary layer as turbulent eddies generated near the Earth's surface reach high enough for condensation to occur.

A cumuliform forms in the free atmosphere if a parcel of air, upon saturation, is warmer than the surrounding ambient atmosphere. Because the parcel is warmer than its surroundings, it accelerates upward, creating the saturated turbulent bubble characteristic of a cumuliform cloud. Cumuliform clouds extending no deeper than the lower troposphere are called cumulus humulis when they are randomly distributed and stratocumulus when they are organized into lines. Cumulus congestus clouds extend into the middle troposphere, while cumulonimbus are the deep precipitating cumuliform clouds that extend throughout the troposphere. Cumulonimbus clouds develop from cumulus humulis and cumulus congestus.

The tropopause, the top of the troposphere, corresponds to the level at which the general decrease in temperature within the troposphere ceases and is replaced by an essentially isothermal layer. In the tropics and subtropics the tropopause is high, often reaching to about 18 kilometres as a result of the vigorous vertical mixing of the atmosphere by thunderstorms. In contrast, in polar regions, where such deep atmospheric turbulence is much less fre-

quent, the tropopause is often as low as eight kilometres. Temperatures at the tropopause height range from −80° C in the tropics to −50° C in the polar regions.

Balloon-borne radiosondes

Meteorologic observations. Radiosondes borne aloft by helium balloons are the primary observation platforms above the boundary layer but within the limits of the troposphere. They measure atmospheric temperature, dew point temperature, and barometric pressure. Since their position is tracked by radar, wind speed and direction as a function of height are also routinely provided. (Radiosondes are sometimes called rawindsondes for this reason.) Meteorologists have been able to produce a long-period data record of the state of the troposphere with radiosonde measurements. The instruments are sent up twice a day simultaneously around the world. Meteorologic observations from radiosondes also are used to initialize the numerical weather prediction models employed for forecasting day-to-day weather conditions.

Another more advanced type of remote-sensing system known as a profiler has been developed to provide almost continuous measurements of wind and, sometimes less accurately, of moisture and temperature throughout the lowest 10 kilometres of the atmosphere. Wind speeds are estimated by upward-directed Doppler radars, while temperature and moisture profiles are evaluated by vertically pointing radiometers that measure electromagnetic emissions of selected wavelengths from various heights in the troposphere. Used in conjunction with satellite-based passive temperature and moisture radiometric soundings and active lidar wind measurements, profilers may eventually render radiosonde soundings obsolete.

Observations from aircraft

Aircraft also provide detailed information about the structure of the atmosphere, particularly during field experiments. Such airplanes as the NOAA P-3 are equipped with a wide array of instruments, including Doppler radar, turbulence sensors, and devices for the in situ measurement of the water and ice content of clouds and their structure. The NOAA P-3 has been used to fly through hurricanes and other deep precipitating cloud systems.

The upper regions of the atmosphere. *Characteristics.* The stratosphere is located above the troposphere and extends up to about 50 kilometres. Above the isothermal layer in the lower stratosphere, temperature increases with height. Temperatures as high as 0° C are observed near the top of the stratosphere. This temperature increase is a result of solar heating as ultraviolet radiation in the wavelength range of 0.200 to 0.242 micrometre dissociates diatomic oxygen. The resultant attachment of single oxygen atoms to O_2 produces ozone. The observed increase in temperature with height gives rise to strong thermodynamic stability (*i.e.,* large and positive values of *Ri*), with little turbulence and vertical mixing. Because of the warm temperatures and very dry air, the stratosphere is almost cloud free.

The ozone layer

Ozone is produced mainly at tropical and mid-latitudes in the stratosphere; maximum destruction of this form of oxygen occurs in the same locations through catalytic cycles of the nitrogen oxides. Ozone is also transported downward—primarily in the vicinity of the polar front—and poleward, resulting in maximum vertical content in the chemically inactive polar region.

The top of the stratosphere is capped by the stratopause. Above this height, which occurs around levels near 45–50 kilometres and pressures of one millibar, is the mesosphere, where temperatures again decrease with height. Vertical air currents, unlike the stratosphere, are not strongly inhibited in this layer. Ice crystal clouds, known as noctilucent clouds, occasionally form in the upper mesosphere.

Above a height of about 85 to 90 kilometres—the mesopause—temperature again increases with height. The layer above this is the thermosphere, where temperatures range from about 500 K during periods when the Sun is inactive to 2,000 K when it becomes active. The thermopause, defined as the level of transition to a more or less isothermal temperature profile at the top of the thermosphere, occurs at heights of about 250 kilometres during quiet Sun periods to almost 500 kilometres when the Sun is active. Above 500 kilometres, molecular collisions are infrequent enough that temperature is difficult to define.

The part of the thermosphere where charged particles, or ions, are abundant is the ionosphere. The ions result from the removal of electrons from gases by ultraviolet solar radiation. The ionosphere, which extends from about 80 to 300 kilometres in altitude, is an electrically conducting layer from which radio signals can be reflected. Shortwave radio transmissions that can reach around the world take advantage of the ability of the ionospheric layers to reflect certain wavelengths of electromagnetic radiation.

Above about 500 kilometres the motion of ions is strongly constrained by the presence of the Earth's magnetic field. This region of the atmosphere, called the magnetosphere, is compressed by the solar wind on the daylight side of the Earth and stretched outward in a long tail on its night side. The colourful auroral displays often seen in polar latitudes are associated with the generation by solar energy outbursts of high-energy particles in the magnetosphere, which are subsequently injected into the lower ionosphere.

The magnetospheres and auroras

The layer above 500 kilometres is also referred to as the exosphere. In this region at least half of the upward-moving molecules do not collide with one another but rather follow long ballistic trajectories, exciting the atmosphere completely if their escape velocities are high enough. The rate of the molecules lost through the exosphere is critical in determining whether or not the Earth, or other planetary body, retains an atmosphere.

The terrestrial atmosphere is also segmented into a lower layer called the homosphere in which turbulent mixing dominates molecular diffusion of gases. In this region, which occurs below 100 kilometres or so, atmospheric composition tends to be independent of height. Above 100 kilometres, however, in the region designated the heterosphere, the lighter gases are concentrated in the highest layers. Above 1,000 kilometres helium and hydrogen are the dominant species. The relatively heavy gas diatomic nitrogen drops off rapidly with height, and only traces remain at 500 kilometres. This decrease in concentration of heavier gases with height is largest during periods of low solar activity when the temperatures in the heterosphere are relatively low. The transition zone at a height of around 100 kilometres between the homosphere and heterosphere is called the turbopause.

Atmospheric tides

Tides occur in the atmosphere, with the greatest magnitudes in the upper atmosphere, due to direct diurnal heating of the air as the Earth rotates and due to solar and lunar gravitational effects. In contrast to the ocean, the generation of tides by heating is much more important than the gravitational effect.

Observations of the upper atmosphere. Above the usual maximum height of radiosonde measurements (above about 17 kilometres and 100 millibars), rocketsondes, along with rocket-borne grenade and falling-sphere experiments, have been used to monitor thermal structure. Since these measurements are much less frequent than radiosonde observations, however, less is known about the meteorologic phenomena above the tropopause than at lower altitudes. Satellite radiometric soundings also have been used to provide temperature structure down to 60 kilometres or so, though they have less vertical and temporal resolution than in situ measurements. In addition, investigators make use of ground-based radar to measure atmospheric characteristics in the upper atmosphere.

Satellite radiometric soundings

STUDY OF THE HORIZONTAL STRUCTURE OF THE ATMOSPHERE

The motions of the atmosphere and the physical processes involved in air flow are studied by dynamicists. Since the 1960s an increasing number of such investigators in dynamic meteorology have adopted computer models as a basic research tool. Computer models of general global circulation and of various small-scale motion systems (*e.g.,* hurricanes and squall lines) have enabled them to better understand the structure and physics of the atmosphere.

Asymmetrical distribution of solar heating and its effects. The primary driving force on the Earth's atmosphere is the amount and distribution of solar radiation that impinges on the planet. The orbit of the Earth around the Sun is an ellipse with an apogee of 1.47×10^8

kilometres in early January and a perigee of 1.52×10^8 kilometres in early July. The time between the autumnal equinox and the following vernal equinox in the Northern Hemisphere (about September 22 to March 21) is about one week shorter than the rest of the year because of the Earth's elliptical orbit. This results in shorter winters in the Northern Hemisphere than south of the Equator.

The Earth rotates every 24 hours around an axis that is tilted at an angle of 23°30′ with respect to the plane of its orbit. As a result of this tilt, sunshine is more direct on a flat surface at a given latitude during summer than it is during winter in either the Northern or Southern Hemisphere. Poleward of latitude 66°30′, the tilt of the Earth is such that for at least one complete day (at 66°30′) and as long as six months (90°) the Sun is above the horizon during summer and below the horizon during winter.

This asymmetrical distribution of solar heating has the following effect. In winter the high latitudes become very cold in the troposphere as a result of the long nights. In summer the troposphere warms significantly at high latitudes as a result of the long hours of daylight. Yet, because of the oblique angle of the sunlight, the temperatures remain, in general, relatively cool compared to regions in the mid-latitudes during summer. Equatorward of 30° or so, however, substantial and similar radiational heating from the Sun occurs during both winter and summer. The tropical troposphere, therefore, has comparatively little variation in temperature during the entire year.

In the troposphere the demarcation between the cold polar air and the warmer tropical air is usually well defined by the polar front—poleward of the front the air is of polar origin, equatorward it is of tropical origin. The colder polar air is denser than the tropical air, with more than a 30 percent difference in densities at the surface possible for extreme wintertime contrasts. During winter the polar front is generally located at lower latitudes and is stronger than in summer.

The region of greatest solar heating at the surface in the humid tropics results in areas of deep cumulonimbus convection. These cumulonimbus clouds occur because, upon condensation, the clouds are warmer than the surrounding ambient atmosphere. These clouds transport water vapour, sensible heat, and the Earth's rotational momentum to the upper portion of the troposphere.

Atmospheric circulation patterns. Because motion upward into the stratosphere is inhibited by a very stable thermal stratification, the air transported upward by the convection diverges poleward in the upper troposphere. This divergence aloft results in a minimum of pressure at the surface, which is called the equatorial trough. As the air is transported poleward, it is deflected toward the right in the Northern Hemisphere and toward the left in the Southern Hemisphere since it tends to retain the angular momentum of the near-equatorial region. At low latitudes the angular momentum is large because of the Earth's rotation.

Upon reaching about latitude 30° poleward of its region of origin, the air is traveling primarily toward the east. Since all upward motion is constrained by the stratosphere, the air must descend. As the air descends, the resultant compressional warming creates vast regions of strong thermodynamic stability within the troposphere. The sparse precipitation in these regions, a result of stabilization and subsidence, is associated with the great arid regions of the world, such as the Sahara, Atacama, Kalahari, and Sonoran deserts. The accumulation of air due to convergence in the upper troposphere causes deep high-pressure systems called subtropical ridges to occur in these regions.

Upon reaching the lower troposphere, the presence of the Earth's surface requires that the air diverge, with some air moving poleward and the rest equatorward. In either direction the air is deflected to the right in the Northern Hemisphere and to the left in the Southern Hemisphere.

Horizontal deflection of air

The tendency of an air parcel to conserve its momentum causes this horizontal deflection. Such deflection occurs because, according to Newton's first principle, a parcel in motion in a certain direction retains the same motion unless acted on by an exterior force. With respect to the rotating Earth, therefore, a moving parcel that is conserving its momentum (*i.e.*, not acted on by an exterior force) will appear to be deflected with respect to fixed points on the rotating planet. As seen from a fixed point in space, a parcel that is conserving its momentum would be moving in a straight line. This apparent force on air motion is the Coriolis effect. Because of this force, air rotates clockwise around large-scale, low-pressure systems but counterclockwise around large-scale, high-pressure systems in the Northern Hemisphere. The flow direction is reversed in the Southern Hemisphere. In the equatorward-moving flow, this deflection results in northeast winds north of the Equator and southeast winds south of the Equator. These low-level winds are the trade winds. The low-level convergence region of the northeast and southeast trade winds from the two hemispheres is the intertropical convergence zone (ITCZ). The ITCZ corresponds to the equatorial trough and is the mechanism that helps to generate the deep thunderstorms in this pressure trough.

The circulation of ascent in the equatorial trough, poleward movement in the upper troposphere, descent in the subtropical ridges, and equatorward movement in the trade winds is a "direct heat engine" known as the Hadley cell. It is a persistent circulation feature whereby heat from the latitudes of greatest solar insolation is transported to the latitudes of the subtropical ridges. The geographic location of the Hadley circulation moves north and south with the seasons.

Poleward of the subtropical ridges in the lower troposphere, southwesterlies generally occur in the Northern Hemisphere and northwesterlies in the Southern because of the tendency for air motions to conserve absolute angular momentum. Since warm air is being moved poleward at low levels, however, the wind flow is no longer associated with a direct heat engine. The heat that originated in the equatorial trough is consequentially transported further poleward by large, horizontal low-pressure eddies called extratropical cyclones. These extratropical cyclones develop on the polar front when a sufficiently large horizontal gradient of temperature in the lower troposphere develops across the front. The intensity of this temperature gradient is referred to as the baroclinicity of the front.

Developmental stages of extratropical cyclones

Extratropical cyclones are found to have three stages of development: (1) the developing stage in which an undulating wave forms along the front; (2) the mature stage in which sinking cold air sweeps equatorward west of the surface low and ascending warm air moves poleward east of the cyclone; and (3) the occluded stage in which the warm air has become entrained within and moved above the air of polar origin and cut off from the source region of the tropical air. Cyclones that evolve no further than the developing stage are called wave cyclones, while extratropical lows that reach the mature and occluded stages are baroclinically unstable waves. Extratropical storm development is termed cyclogenesis. Surface pressure falls of more than about 24 millibars per day that occasionally occur with rapid extratropical cyclone development are referred to as explosive cyclogenesis and are often associated with major winter storms. Theoretical analysis has shown that the occurrence of baroclinically unstable waves is directly proportional to the magnitude of the temperature gradient, with maximum growth for wavelengths of 3,000 to 5,000 kilometres.

Fronts

Cold fronts occur at the leading edge of polar air moving toward the Equator, while warm fronts are defined at the equatorward surface position of the polar air as it retreats poleward east of the extratropical cyclone. The equatorward-moving air behind the cold front occurs in pools of cold, dense, polar and arctic surface high-pressure systems. Arctic highs are defined to distinguish air of an origin even deeper within the high latitudes than the polar highs. When the polar air is neither retreating nor advancing, the polar front is said to be a stationary front. In the occluded stage, where the cold air west of the surface low-pressure centre advances more rapidly eastward around the cyclonic circulation than the cold air east of the centre moves poleward, the warm, less dense tropical air is forced aloft. The resultant frontal intersection is an occluded front. Fronts of all types always move in the direction toward which the colder air is moving.

Clouds and often precipitation occur poleward of the warm and stationary fronts whenever the poleward-moving, less dense tropical air north of subtropical ridges reaches the latitude of the polar front and is forced upward over the colder air near the surface. Such fronts are defined as active fronts, and rain and snow from them constitute a major part of the precipitation received in middle and high latitudes, particularly in winter.

The polar front slopes toward the colder air with height. This occurs because cold air, being denser, tends to undercut the warmer air of tropical origin. Since the cold air is denser, pressure decreases more rapidly with height poleward of the polar front than on the warmer side. In the middle and upper troposphere, the resultant large horizontal pressure gradient between the polar and tropical air creates strong westerly winds as air circulates around the region of low pressure in the higher latitudes at these heights. The centre of this low-pressure region is the circumpolar vortex. The region of strongest winds, which occurs at the juncture of the tropical and polar air masses, is called the jet stream. Since the temperature contrast between the tropics and the high latitudes is greatest in winter, the jet stream is stronger during that season. In addition, since the mid-latitudes also become colder during winter while tropical temperatures remain relatively unchanged, the westerly jet stream tends to move equatorward during the colder season.

Seasonal variations of the jet stream

The jet stream reaches its greatest velocities at the tropopause. Above that level, lower tropopauses in the polar region than in the tropics result in a reversal of the horizontal temperature gradient in the stratosphere from that found in the troposphere, with warmer temperatures at high latitudes. This causes a weakening of the westerlies with height. At intervals of 20 to 40 months, with a mean of 26 months, a reversal of wind direction occurs at low latitudes in the stratosphere, so that easterly flow develops. This feature is called the quasi-biennial oscillation. A phenomena known as sudden stratospheric warming, apparently a result of strong downward motion, also occurs in late winter and spring at high latitudes and can significantly influence the chemical balance of ozone and other reactive gases in the stratosphere.

Preferred geographic locations exist for the development, movement, and decay of extratropical cyclones and for the presence of centres of the subtropical ridge. During winter in the middle and high latitudes, continents tend to become lower tropospheric high-pressure reservoirs of cold air as heat is radiated out to space during the long nights. In contrast, oceans lose heat less rapidly due to several factors: the large thermal inertia of water; its ability to overturn as the surface cools and become negatively buoyant; and the existence of such ocean currents as the Gulf Stream and the Kuroshio that transport heat from lower latitudes poleward. The lower troposphere over the warmer oceanic areas thus tends to be regions of relative low pressure. As a result of this juxtaposition of cold and warm air, the east sides of continents and the western fringes of oceans in the middle and high latitudes are preferred locations for extratropical storm development. Over the Asian continent, in particular, the cold high-pressure system is sufficiently permanent that a persistent offshore flow called the winter monsoon occurs.

Monsoons and related phenomena

An inverse type of flow develops in summer as the continents heat more than adjacent oceanic areas. Continental areas become regions of relative low pressure, while high pressure in the lower troposphere becomes more prevalent offshore. Persistent lower tropospheric onshore flow that develops over large landmasses as a result of the heating is referred to as the summer monsoon. The leading edge of this monsoon is associated with a trough of low pressure—the monsoon trough. Tropical moisture brought onshore by the monsoon often results in copious rainfall.

The subtropical ridge is segmented into surface high-pressure cells because of the continental effect. In the subtropics large landmasses tend to be relative centres of low pressure as a result of the strong solar heating. Persistent high-pressure cells, therefore, occur over the oceans. The oval shape to these cells causes a different thermal structure in the lower troposphere on their eastern and western sides. On the east, subsidence from the Hadley circulation is enhanced as a result of the tendency for equatorward-moving air to sink in order to preserve its angular momentum on the rotating Earth. On the western sides of the high-pressure cell, poleward-moving air must ascend in order to preserve its momentum. As a result of the enhanced descent in the eastern oceans, landmasses adjacent to these areas are generally deserts such as those in northwestern and southwestern Africa. In spite of being under the descending branch of the Hadley cell, western continental fringes of the subtropical oceans, by contrast, are more likely to have precipitation because the stabilization effect of the subsidence portion of the Hadley cell is minimized by the upward vertical velocity associated with the western side of circular subtropical high-pressure cells.

The aridity found along the west coasts of continents in subtropical latitudes is further enhanced by the influence of the equatorward surface flow around the high-pressure cells on the ocean currents. This flow exerts a shearing stress on the ocean surface, which results in the deflection of the layer of water above the oceanic thermocline to the right in the Northern Hemisphere and to the left in the Southern Hemisphere. This deflection is a result of the tendency for the water to conserve its angular momentum and, therefore, to move westward when displaced toward the Equator. Cold, lower-level water from below the thermocline rises to the surface to replace this offshore ocean flow. These areas of cold, coastal surface waters result in enhanced atmospheric stability in the lower troposphere and an even further reduction in the likelihood for precipitation, though fogs and low stratus clouds are common.

The effects of elevated terrain on climatic conditions

North–south mountain barriers and large massifs (*e.g.*, the Rocky Mountains and the Tibetan Highlands) also influence atmospheric flow. By imposing a barrier to the general westerly flow in mid-latitudes, air tends to be blocked and transported poleward west of the terrain and equatorward east of the obstacle. Air that is forced up the barrier often is sufficiently moist to produce considerable precipitation on windward mountain slopes, while subsidence on the lee slopes produces more arid conditions. The elevated terrain affects the atmosphere as if it were an anticyclone, with the result that warm air is transported further toward the pole west of the terrain. It is also difficult for cold air in the interior to move westward of the terrain; therefore, relatively mild weather exists for the latitude, as, for example, along the west coast of North America. By contrast, east–west mountain barriers (*e.g.*, the Alps) offer little impediment to the general westerly flow, resulting in maritime conditions extending far inland.

A major focus of weather forecasting in the middle and high latitudes is to predict the movement and development of extratropical cyclones, polar and arctic highs, and the location and intensity of subtropical ridges.

STUDY OF CLOUD PROCESSES

Elements of cloud formation. The formation of cloud droplets and cloud ice crystals occurs associated with suspended aerosols of natural and anthropogenic origin, which are ubiquitous in the Earth's atmosphere. In the absence of such aerosols, relative humidities much greater than 100 percent with respect to a flat surface are required for water vapour to condense spontaneously into liquid water or to deposit into ice. The development of clouds without aerosols, which occurs only in a controlled laboratory environment, is referred to as homogeneous nucleation. In the atmosphere, aerosols serve as initiation sites for the condensation or deposition of water vapour. By having a discrete size, they reduce the amount of supersaturation required for water vapour to change its phase. (Air containing water vapour with a relative humidity greater than 100 percent with respect to a flat surface is said to be supersaturated.) Aerosols that are effective as embryonic sites for the conversion of water vapour to liquid water are called cloud condensation nuclei. The larger the aerosol and the greater its solubility, the lower is the supersaturation required for the aerosol to serve as a cloud condensation nuclei. Cloud condensation nuclei in the atmosphere become effective at supersaturations of about 0.1 to 1 percent. The concentration of such nuclei in the

Cloud condensation nuclei and ice nuclei

lower troposphere at a supersaturation of 1 percent range from roughly 100 per cubic centimetre in oceanic air to 500 per cubic centimetre in a continental atmosphere.

Aerosols that are effective for the conversion of water vapour to ice crystals are called ice nuclei. In contrast to cloud condensation nuclei, the most effective ice nuclei are hydrophobic having molecular spacings and a crystallographic structure close to that of ice.

While cloud condensation nuclei are always readily available in the atmosphere, ice nuclei are often deficient. (Examples of condensation nuclei include particles of sodium chloride [salt] and ammonium sulfate, while the clay mineral kaolinite serves as ice nuclei.) Consequently, liquid water that is lifted and cooled below 0° C can often remain liquid at subfreezing temperatures because of the absence of effective ice nuclei. Except for ice crystals that are effective at 0° C, all other ice nuclei become effective only at temperatures lower than freezing. Liquid water at temperatures less than 0° C is referred to as supercooled water. In the absence of any ice nuclei, freezing of supercooled water droplets that measure a few micrometres in radius requires temperatures at or lower than −39° C (a process called homogeneous ice nucleation). When ice nuclei are present, heterogeneous ice nucleation can occur at warmer temperatures.

Ice nuclei are of three types: deposition, contact, and freezing. Deposition nuclei act analogously to condensation nuclei in that water vapour deposits directly as ice crystals on the aerosol. Contact and freezing nuclei, in contrast, are associated with the conversion of supercooled water to ice. Contact nuclei act to convert liquid water to ice by the touching of the contact nuclei aerosol to the supercooled water droplet. Freezing nuclei are absorbed into the liquid water and convert the supercooled water to ice from the inside out.

Precipitation processes in clouds. The evolution of clouds after cloud droplets or ice crystals have formed depends on which phase of water occurs. A cloud in which only liquid water occurs (even at temperatures less than 0° C) is called a warm cloud, and precipitation emanating from such a cloud is said to be the result of warm-cloud processes. In such a cloud, the growth of liquid water from a cloud droplet to a raindrop occurs first as continued condensational growth due to additional water vapour condensing in a supersaturated atmosphere. This process is effective, however, only until the droplet attains a radius of about 10 micrometres. Above this size, further increases in its radius by condensational growth are very slow, since the mass of the droplet increases as the cube of its radius. Subsequent growth thus occurs only when the cloud droplets develop at slightly different rates due to spatial variations in the initial aerosol sizes and solubilities and to magnitudes of supersaturation. Cloud droplets of different sizes fall at different velocities such that cloud droplets of different radii collide. If the collision is strong enough to overcome the surface tension between the two colliding droplets, coalescence occurs with a new, larger single droplet resulting. This process of cloud droplet growth is referred to as collision-coalescence. Warm-cloud rain results when the droplets attain a sufficient size to fall to the ground. Such a raindrop (perhaps about one millimetre in radius) contains on the order of 1,000,000 cloud droplets of 10-micrometre radius. This type of precipitation is common from shallow cumulus clouds over tropical oceans where the concentration of cloud condensation nuclei is small enough that there is only limited competition for the available water vapour.

Warm-cloud processes

A cloud that contains ice crystals is termed a cold cloud, and precipitation resulting from such a cloud is said to be due to cold-cloud processes. In this type of cloud, ice crystals can grow either by deposition directly from water vapour that is supersaturated with respect to ice or as a result of the evaporation of supercooled water and subsequent deposition onto the ice crystal. Because the saturation vapour pressure of liquid water is greater than or equal to the saturation vapour pressure of ice, ice crystals grow at the expense of the liquid water. For example, air that is saturated with respect to liquid water is supersaturated with respect to ice by 10 percent at −10° C and by 21 percent at −20° C. This results in a rapid conversion of liquid water to ice. In a cloud with numerous supercooled cloud droplets, such a substantial and rapid change of phase permits large ice crystals of snowflake size to develop quickly from tiny crystals by depositional growth alone. Clouds that are converted only to ice crystals are known as glaciated clouds. Ice crystals that grow by deposition have much lower densities than solid ice because of the air pockets occurring within the volume of the crystal.

Cold-cloud processes

The specific form of the ice crystals that form depends on the temperature and the degree of supersaturation with respect to ice. At −14° C and a relatively large supersaturation with respect to liquid water, for example, dendritic ice crystals form. This type of ice crystal has growth at the end of radial arms on one or more planes of the crystal. At −40° C and a supersaturation with respect to liquid water of close to zero, hollow ice columns form.

Ice crystals also can grow to precipitation size through aggregation or by riming. Aggregation occurs when the arms of the ice crystals interlock, resulting in a clump. This collection of intermingled ice crystals occasionally can attain sizes of several centimetres in diameter. In riming, supercooled water freezes directly onto ice crystals, causing them to grow; the accumulation of dense ice on the crystals increases their fall velocity. When the riming is substantial, the crystal form of the snowflake is lost and a more or less spherical precipitation-sized particle called graupel results. In cumulonimbus clouds wherein the graupel is repeatedly wetted and then injected back toward high altitudes as a result of strong updrafts, very large graupel, or hail, results.

Aggregation and riming

Frozen precipitation falling to levels much warmer than 0° C reaches the surface as rain. Such cold-cloud rain at the ground is distinguished from warm-cloud rain by its larger size. Melted hailstones, in particular, make a large radius impact when they strike the ground. Cold-cloud rain may refreeze if a layer of subfreezing air exists near the surface. If the freezing occurs in the free atmosphere, sleet or ice pellets is produced. When the freezing occurs only at impact on the ground, freezing rain results.

Procedures of cloud physics research. The study of cloud formation and precipitation processes in clouds is conducted by cloud physicists. The presence of cloud condensation and ice nuclei in air parcels is tested with cloud chambers in which controlled temperatures and relative humidities are specified. In the actual atmosphere, aircraft fly through clouds and collect droplets and ice on collection plates or photograph their presence in the airstream. In the past, identification of the different sizes of droplets and the various types of ice crystals was performed subjectively by researchers, though computer-image assessment procedures now are automating this analysis. At the ground, rainfall impaction molds and snow crystal impressions are made. Hailstones also are collected because an analysis of their structure helps to define the ambient environment in which they formed. Chemical analyses of the cloud droplets, ice crystals, and precipitation are also frequently carried out in order to identify natural and man-made pollutants within the different forms of water.

Cloud chamber studies

STUDY OF CLIMATE AND CLIMATIC CHANGE

As noted earlier, climate is generally conceived of as the expected weather conditions for specific geographic locations. It is defined in terms of averages and as standard deviations around the average.

Local determining factors. Through extensive observations climatologists have found that the primary influences on climate include latitude, degree of continentality, character of the surface, and elevation of a location. Latitude directly influences the intensity of sunlight reaching the ground. In the polar regions, for example, little or no solar radiation reaches the surface in midwinter, while at the Equator the annual variation of solar intensity is small.

The degree of continentality determines the extent to which nearby marine areas moderate temperature variations. Land areas heat and cool relatively rapidly because of the low thermal conductivity of ground surfaces. In contrast, water bodies have a much larger thermal inertia

as a result of the ability of sunlight to penetrate to significant depths and the capability of water to mix vertically. Consequently, diurnal and annual temperature variations over marine areas tend to be much smaller than over continental sites in the same latitude. Observation sites on land but near water bodies tend to have climates that are moderated by the marine influence, particularly when the average wind flow is from the water toward the land. Areas with a large continentality are generally drier than marine areas because such regions are distant from the oceanic source of water vapour to the atmosphere.

The character of the surface also directly influences local climate. A heavily vegetated surface tends to have smaller temperature variations than bare ground. Ground without vegetation converts a relatively large fraction of solar radiation into sensible heating of the air. In contrast, solar radiation on vegetation produces evapotranspiration, which reduces the magnitude of sensible heating. At night bare soil effectively radiates heat into space, thereby producing substantial cooling. Vegetation, on the other hand, helps insulate the ground from this heat loss. Snow-covered ground produces a cooler climate than would otherwise occur because of its greater reflection of incident sunlight. Evapotranspiration from vegetation moistens the atmosphere, making precipitation more likely.

Elevation directly influences climate as temperature normally decreases with height in the troposphere. Elevated areas thus tend to have cooler temperatures than lower sites at the same latitude. Also, solar intensity during the day and long-wave radiative loss at night are larger at higher altitudes because the overlying atmosphere is thinner. This often results in larger diurnal variations in temperature at higher elevations than occur closer to sea level.

The relation of precipitation patterns to general circulation features

The climatic distribution of precipitation can be related to the major general circulation features of the Earth. Air ascends on the average in the intertropical convergence zone and along the polar front. The intertropical convergence zone, which separates air of northern and southern hemispheric origin, is associated with deep thunderstorms, while migratory extratropical cyclones propagate along the polar front, producing large-scale organized areas of precipitation. Locations close to these major weather features throughout the year tend to have substantial precipitation evenly distributed over the seasons.

Descending motion (subsidence) is associated with the subtropical ridge and the circumpolar arctic high-pressure region. Little precipitation occurs in those areas that are dominated by these weather features. Arid climatic regions such as the Sahara, Kalahari, and Sonoran deserts result directly from the persistence of subtropical ridges over these areas throughout the year. Regions that are influenced by the subtropical ridge only during the summer (*e.g.*, southern California, southwestern Australia, and the Mediterranean area) but are often in the vicinity of the polar front during winter have a dry summer/wet winter climate pattern. The Antarctic continent is very dry because of its location relative to the subsidence of the circumpolar Antarctic high-pressure region.

Climatic change and its causes. Humans have adjusted agricultural and other activities to the current climatic configuration of the Earth. Climatic conditions, however, change with time, as, for example, from the apparent warm, humid global conditions of the Carboniferous Period to the widespread continental glaciation of the Pleistocene Epoch. Using fossils and other geologic evidence (*e.g.*, erosional landforms, shoreline features, and glacial deposits), paleoclimatologists have demonstrated that the periodic occurrence of extensive glaciation separated by long periods of a warm global climate is a recurrent characteristic of the Earth. The causes of these climatic changes have been attributed to a variety of mechanisms, including increased volcanic emissions that have been associated with the blocking of sunlight and the resultant cooling at the surface. Periodic reductions in solar output also have been suggested as the cause of global cooling.

Findings of paleoclimatological research

The movement of the continents over the Earth's surface over long time periods is thought to have caused different global climatic patterns. This migration of the landmasses, known as continental drift, has been invoked to explain geologic evidence of tropical fauna in Antarctica and of glaciers at low altitudes in Africa.

Variations over time of the obliquity of the Earth's axis with respect to its orbital plane, the eccentricity of the orbit, and the precession of the axis directly influence the distribution of solar radiation over the planet and therefore the climate. The obliquity of the Earth varies between 24°36′ and 21°39′ from its current value of 23°30′ over a period of approximately 40,000 years. The eccentricity ranges between about 0 to 0.05 from its current value of 0.016 over a time period of about 92,000 years, while the precession of the axis requires from 16,000 to 26,000 years to make a complete circle. The most pronounced difference between winter and summer seasons occurs with a large obliquity and a large eccentricity such that winter occurs when the Earth is farthest from the Sun.

Effects of human activities on climate

Over the last few hundred years, humankind has been directly influencing global and local climate. The development of urban areas has created different ground characteristics that have resulted in urban heat islands in which cities are warmer, particularly at night, than the surrounding countryside. The input of carbon dioxide (CO_2) into the atmosphere through industrial activities has been suggested to be associated with warming near the surface as additional long-wave radiation emitted at the surface is absorbed by the CO_2 and radiated back toward the surface. In the period 1958–75, for example, the average CO_2 level of the atmosphere increased at a rate of about 1.7 parts per million per year. There is concern that by the year 2100 the enhanced CO_2 level resulting from industrial activity will increase the average global temperatures by as much as 5° C, with the greatest impact at high altitudes.

Aerosols are also released into the atmosphere by industrial and other human activities. Climatologists have suggested that anthropogenic-generated aerosols could alter the Earth's radiation budget, perhaps even counteracting the warming effect of CO_2. The ability of additional aerosols to heat or to cool the Earth's atmosphere depends on their vertical and horizontal distribution, and their concentration, size, and chemistry.

The addition to the atmosphere of anthropogenic aerosols, which serve as additional cloud condensation and ice nuclei, also could alter the percentage of the Earth covered by clouds. Increased concentrations of cloud condensation nuclei, for instance, would reduce the average droplet size within a cloud, making the droplets more colloidally stable and thus less likely to precipitate. Such clouds are likely to persist longer, resulting in enhanced reflection of sunlight during the day (*i.e.*, a cooling effect) but a reduction of long-wave radiational cooling at night if the clouds are in the low to middle troposphere. The net effect on global climate remains unclear.

PRACTICAL APPLICATIONS

Weather forecasting. Among the various applications of the atmospheric sciences, weather forecasting is of considerable immediate importance. A necessary condition to making accurate predictions of future weather conditions is an adequate characterization of the pattern of winds, temperatures, moisture, and pressure that currently prevail in the troposphere across the globe. Data for such a meteorologic analysis is obtained by the aforementioned balloon-borne radiosondes that are released twice daily (at 0000 and 1200 Greenwich Mean Time) around the world. Global analyses of temperature, moisture, and winds are routinely prepared for the surface as well as for the standard pressure levels of 850, 700, 500, 300, and 200 millibars. Weather information is shared worldwide via the Global Telecommunications System (GTS).

Numerical weather prediction models

The principal tools used in weather forecasting include numerical weather prediction models. Employing mathematical representations of the physical conservation laws of motion, heat, mass, and moisture in the form of nonlinear, partial differential equations, synoptic meteorologists are able to approximate relations for solution on a three-dimensional grid mesh and to integrate them forward in time. The grid mesh corresponds to the domain of interest, which can represent the entire Earth (a global model) or a local region such as North America (a limited

area model). Exact solutions of the nonlinear equations are, in general, not possible, necessitating their evaluation on a grid. Large, high-speed supercomputers (*e.g.*, the CYBER 205 systems of the United States National Meteorological Center and of the British Meteorological Office, and the CRAY systems of the U.S. National Center for Atmospheric Research and of the European Centre for Medium Range Forecasting) are required to perform the vast number of calculations associated with the integration of atmospheric models forward in time. The use of such supercomputer-integrated models appears to have made it possible to accurately predict pressure fields and temperature anomalies and, to a lesser extent, precipitation at least five to seven days in advance. Public forecasts of weather beyond 12 hours are almost exclusively based on the predictions and statistically derived products of these complex models.

Reliance on weather satellite, radar, and surface observations

For time periods of less than 12 hours, radar, satellite, and surface observations are used extensively to characterize details of local weather for its short-term extrapolation into the future. In a form of prediction known as nowcasting, these sources of information are also used to fill in spatial variations in weather that are both unobservable and incapable of being predicted in large-scale weather forecasting models with horizontal grid-mesh distances of 80 kilometres or more. Tracking a hurricane and following the evolution of a squall line are examples in which high-resolution observation platforms are used to monitor weather systems. Numerical prediction models with smaller grid meshes (on the order of one to 20 kilometres) have been applied over several-hundred-kilometre by several-hundred-kilometre areas to predict weather 12 to 18 hours ahead. This scale of atmospheric motion is defined as the mesoscale.

Weather modification. In addition to predicting weather, man is seeking to deliberately alter atmospheric conditions. Most past and present efforts to modify the weather involve attempts to augment precipitation, suppress hail and fog, and weaken hurricanes.

Cloud seeding

Precipitation increases have been sought by both static and dynamic cloud seeding. One technique of static cloud modification involves introducing modest amounts of silver iodide or solid carbon dioxide (Dry Ice), from aircraft or ground-based rockets, into subfreezing levels of a liquid water cloud in which natural ice nuclei are deficient. Silver iodide, which has a crystal lattice structure similar to that of ice, provides artificial ice nuclei, while Dry Ice directly freezes the liquid water. Because snow crystals grow rapidly in a subfreezing liquid water environment, precipitation results. This mechanism of cloud seeding is used in stratiform clouds upwind of mountainous areas to increase snowpack. It also is used to dissipate subfreezing fog at airports through the conversion of the fog droplets to snow crystals, which precipitate to the ground.

Dynamic cloud seeding involves the massive injection of silver iodide into a subfreezing layer of a growing cumulus cloud composed of water droplets. The rapid conversion of the supercooled water to ice—and associated release of the latent heat of fusion—causes enhanced buoyancy and growth of the cumulus cloud into a larger system than it otherwise would be. Greater precipitation is associated with deeper cumulus clouds. Also, by judiciously seeding adjacent cumulus clouds, it may be possible to promote their merger into large cumulonimbus systems.

The injection of large quantities of silver iodide into potential hailstorms also has been pursued. The hypothesis is that a larger number of ice crystals from which hailstones develop will result in smaller, more numerous hailstones as the competition for the water available among ice particles is greater. Since the hail is smaller it is more likely to melt before reaching the surface and doing damage to crops.

Besides such deliberate attempts at weather modification, humans are inadvertantly altering weather and climate as well. Changes in ground surface characteristics by urbanization, agriculture, and other activities have altered the heat and moisture budget at the surface. The clearing of forests in the Amazon Basin, for example, is believed to be increasing the carbon dioxide content of the atmosphere due to the loss of photosynthetic biomass. The burning of fossil fuels also releases carbon dioxide and is thought to be contributing to the overall increase. Carbon dioxide is an effective absorber and remitter of long-wave radiation released from the Earth, so that higher levels of CO_2 are suspected to cause global warming, possibly to the extent of melting substantial portions of the Greenland and Antarctic ice caps. At the same time, the burning of fossil fuels, along with other industrial activities, releases vast quantities of airborne particulates. Because of the continuous emission of these air pollutants into the tropospheric wind circulation and cumulus cloud venting into the mid- and upper troposphere, more solar radiation may be reflected back into space than would otherwise occur, resulting in a tendency for global cooling. This would certainly be the case if the particulate matter reaches the upper troposphere and stratosphere where its residence time would be of long duration.

Problems associated with the buildup of CO_2 and particulate matter in the atmosphere

Cooling associated with enhanced turbidity from industrial particulates in the atmosphere could balance the warming that is suggested to result from increased CO_2 concentrations. Unfortunately both effects are expected to become much larger during the next century as developing countries industrialize. Should either effect become dominant over the other, a rapid deterioration of the climate could result.

(R.A.Pi.)

BIBLIOGRAPHY

History of the Earth sciences: FRANK DAWSON ADAMS, *The Birth and Development of the Geological Sciences* (1938, reprinted 1954), the best general account for the years prior to 1830; ASIT K. BISWAS, *History of Hydrology* (1970), a factual chronicle of developments since the earliest times; HENRY FAUL and CAROL FAUL, *It Began with a Stone: A History of Geology from the Stone Age to the Age of Plate Tectonics* (1983); A. HALLAM, *A Revolution in the Earth Sciences* (1973), a summary of the historical development of ideas from seafloor spreading to plate tectonics, and *Great Geological Controversies* (1983), an evaluation of celebrated controversies from Neptunism to continental drift; ROBERT MUIR WOOD, *The Dark Side of the Earth: The Battle for the Earth Sciences, 1800–1980* (1985), a history of important controversies; RICHARD J. CHORLEY, ANTONY J. DUNN, and ROBERT P. BECKINSALE, *The History of the Study of Landforms; or, The Development of Geomorphology*, vol. 1, *Geomorphology Before Davis* (1964), an expansive account covering developments to the end of the 19th century; CHARLES C. GILLISPIE, *Genesis and Geology: A Study in the Relations of Scientific Thought, Natural Theology, and Social Opinion in Great Britain, 1790–1850* (1951, reprinted 1969), an analysis of the impact of developments in geology upon Christian beliefs in the decades before Darwin (extensive bibliography); C.P. IDYLL (ed.), *Exploring the Ocean World: A History of Oceanography*, rev. ed. (1972), a symposium treating each of the several branches of oceanography in historical format; JOSEPH NEEDHAM, *Science and Civilisation in China*, vol. 3, *Mathematics and the Sciences of the Heavens and the Earth* (1959), containing a comprehensive and elaborately illustrated account of the history of Earth science in China to around AD 1500; CECIL J. SCHNEER, "The Rise of Historical Geology in the 17th Century," *Isis*, vol. 45, part 3, no. 141, pp. 256–268 (September 1954), an analysis of the points at issue in the fossil controversy; CECIL J. SCHNEER (ed.), *Toward a History of Geology* (1969), 25 essays on the history of geologic thought, mainly of the 18th and 19th centuries; NAPIER SHAW, *Manual of Meteorology*, vol. 1, *Meteorology in History* (1926, reprinted 1932), a rambling but literate and entertaining history of meteorology from the earliest to modern times; EVELYN STOKES, "Fifteenth Century Earth Science," *Earth Sciences Journal*, 1(2):130–148 (1967), an analysis of classical and medieval views of nature, especially those reflected in Caxton's *Mirrour of the World;* PHILIP D. THOMPSON *et al.*, *Weather*, rev. ed. (1980), an introduction to meteorology with much historical material, well illustrated; STEPHEN TOULMIN and JUNE GOODFIELD, *The Discovery of Time* (1965, reprinted 1983), which traces the history of the idea of geologic time; WILLIAM WHEWELL, *History of the Inductive Sciences from the Earliest to the Present Time*, 3rd ed., 3 vol. (1857, reissued 1976)—vol. 2 containing an analysis of uniformitarian and catastrophist views of Earth history; and KARL ALFRED VON ZITTEL, *History of Geology and Palaeontology to the End of the Nineteenth Century* (1901, reissued 1962; originally published in German, 1899, reprinted 1965), best for its history of paleontology.

Geologic sciences: Popular introductions include PETER CATTERMOLE and PATRICK MOORE, *The Story of the Earth* (1985); and FRANK PRESS and RAYMOND SIEVER, *Earth*, 4th ed. (1986). Other readable general accounts are J. DERCOURT and J. PAQUET,

Geology: Principles and Methods, trans. from French (1985), an introduction to the modern Earth: and RICHARD FIFIELD (ed.), *The Making of the Earth* (1985), which reviews recent major breakthroughs in the Earth sciences. (*Study of the composition, structure, and surface features of the Earth*): MARTIN H.P. BOTT, *Interior of the Earth: Its Structure, Constitution, and Evolution*, 2nd ed. (1982), a comprehensive synthesis of the geophysics of the solid Earth; G.C. BROWN and A.E. MUSSETT, *The Inaccessible Earth* (1981), a readable treatment of the geophysics of the Earth's core, mantle, and crust; KENT C. CONDIE, *Plate Tectonics and Crustal Evolution*, 2nd ed. (1982), an advanced but accessible account of the origin and development of the Earth's crust; JOHN SUPPE, *Principles of Structural Geology* (1985), a modern introduction; W. KENNETH HAMBLIN, *The Earth's Dynamic Systems: A Textbook in Physical Geology*, 4th ed. (1985), an introduction to physical geology and tectonics; ARTHUR HOLMES, *Holmes Principles of Physical Geology*, 3rd ed. rev. by DORIS L. HOLMES (1978), a thorough and stimulating coverage of physical geology; AKIHO MIYASHIRO, KEIITI AKI, and A.M. CELÂL ŞENGÖR, *Orogeny* (1982; originally published in Japanese, 1979), which describes orogenic theory based on plate tectonics; CLIFF OLLIER, *Tectonics and Landforms* (1981), an explanation of the interrelationships between these subjects; EDWARD J. TARBUCK and FREDERICK K. LUTGENS, *The Earth: An Introduction to Physical Geology* (1984); SAIYA UYEDA, *The New View of the Earth: Moving Continents and Moving Oceans* (1978), a good introduction to global plate tectonics; P.J. WYLLIE, *The Way the Earth Works* (1976), a readable introduction and overview of the global geology of plate tectonics; PETER FRANCIS, *Volcanoes* (1976), a popular account; ERIC A.K. MIDDLEMOST, *Magmas and Magmatic Rocks: An Introduction to Igneous Petrology* (1985); HAROUN TAZIEFF and JEAN-CHRISTOPHE SABROUX (eds.), *Forecasting Volcanic Events* (1983); and HOWEL WILLIAMS and ALEXANDER R. MCBIRNEY, *Volcanology* (1979), an authoritative and comprehensive account of volcanoes. (*Historical geology, stratigraphy, and paleontology*): P.J. BRENCHLEY (ed.), *Fossils and Climate* (1984), symposium proceedings explaining the influence of past climates on the evolution of life and paleobiogeography; W.S. MCKERROW (ed.), *The Ecology of Fossils* (1978), a comprehensive and illustrated guide; CARL K. SEYFERT and LESLIE A. SIRKIN, *Earth History and Plate Tectonics: An Introduction to Historical Geology*, 2nd ed. (1979), with emphasis on the Phanerozoic period; STEVEN M. STANLEY, *Earth and Life Through Time* (1986), which relates the physical history of the Earth and the history of life; and BRIAN F. WINDLEY, *The Evolving Continents*, 2nd ed. (1984), a detailed, comprehensive synthesis of evidence related to continental evolution throughout geologic time. (*Astrogeology*): MICHAEL H. CARR (ed.), *The Geology of the Terrestrial Planets* (1984), an authoritative account; BILLY P. GLASS, *Introduction to Planetary Geology* (1982), a well-illustrated comprehensive volume on the Earth, the other planets, and the asteroids; and RONALD GREELY, *Planetary Landscapes* (1985), abundant photos of the planets and their satellites and text explaining their geomorphology and geology. (*Applied Earth sciences*): DONALD R. COATES, *Geology and Society* (1985), basic concepts of environmental geology; CHARLES S. HUTCHISON, *Economic Deposits and Their Tectonic Setting* (1983), a comprehensive account of economic geology; and DAVID W. SIMPSON and PAUL G. RICHARDS (eds.), *Earthquake Prediction: An International Review* (1981), which provides a broad spectrum of case histories and an overview of the subject.

(B.F.W.)

Hydrologic sciences: (Hydrology): An excellent introduction is provided by THOMAS DUNNE and LUNA B. LEOPOLD, *Water in Environmental Planning* (1978); while basic concepts in engineering hydrology are covered by RAY K. LINSLEY, JR., MAX A. KOHLER, and JOSEPH L.H. PAULHUS, *Hydrology for Engineers*, 3rd ed. (1982). In more specialized areas, S. LAWRENCE DINGMAN, *Fluvial Hydrology* (1984), treats basic hydraulic principles; R. ALLEN FREEZE and JOHN A. CHERRY, *Groundwater* (1979), deals with subsurface flow processes and water quality; and M.J. KIRKBY (ed.), *Hillslope Hydrology* (1978), is concerned with experimental and modeling studies of catchment processes. Modeling is dealt with further in M.G. ANDERSON and T.P. BURT, *Hydrological Forecasting* (1985). (*Limnology*): Broad introductions are provided by ABRAHAM LERMAN (ed.), *Lakes—Chemistry, Geology, Physics* (1978); and ROBERT G. WETZEL, *Limnology*, 2nd ed. (1983). Some engineering aspects of man-made lakes are covered in B. HENDERSON-SELLERS, *Engineering Limnology* (1984). (*Oceanography*): Excellent introductions are PETER K. WEYL, *Oceanography: An Introduction to the Marine Environment* (1970); KEITH STOWE, *Ocean Science*, 2nd ed. (1983); GEORGE L. PICKARD and WILLIAM J. EMERY, *Descriptive Physical Oceanography: An Introduction*, 4th ed. (1982); and DAVID TOLMAZIN, *Elements of Dynamic Oceanography* (1985). The geology of the ocean basins is the subject of FRANCIS P. SHEPARD, *Geological Oceanography: Evolution of Coasts, Continental Margins and the Deep-Sea Floor* (1977); and the chemistry of the oceans is covered in J.P. RILEY and R. CHESTER, *Introduction to Marine Chemistry* (1971). (*Glaciology*): The classic text is LOUIS LLIBOUTRY, *Traité de glaciologie*, 2 vol. (1964). The physical aspects of ice are covered in W.S.B. PATERSON, *The Physics of Glaciers*, 2nd ed. (1981). Other useful texts are SAMUEL C. COLBECK (ed.), *Dynamics of Snow and Ice Masses* (1980); and D.E. SUGDEN and B.S. JOHN, *Glaciers and Landscape* (1976, reprinted 1979).

(K.J.B.)

Atmospheric sciences: General references on meteorology, climatology, and aeronomy include RALPH E. HUSCHKE (ed.), *Glossary of Meteorology* (1959, reissued 1970); and SMITHSONIAN INSTITUTION, *Smithsonian Meteorological Tables*, 6th rev. ed. prepared by ROBERT J. LIST (1949, reprinted 1968). (*General meteorology and climatology*): STANLEY DAVID GEDZELMAN, *The Science and Wonders of the Atmosphere* (1980); WILLIAM J. KOTSCH, *Weather for the Mariner*, 3rd ed. (1983); GLENN T. TREWARTHA and LYLE H. HORN, *An Introduction to Climate*, 5th ed. (1980); and JOHN M. WALLACE and PETER V. HOBBS, *Atmospheric Science: An Introductory Survey* (1977), more mathematically sophisticated than the above works. (*Dynamic meteorology and climatology*): E. PALMÉN and C.W. NEWTON, *Atmospheric Circulation Systems: Their Structure and Physical Interpretation* (1969); JOHN A. DUTTON, *The Ceaseless Wind: An Introduction to the Theory of Atmospheric Motion* (1976, reissued with corrections, 1986); JAMES R. HOLTON, *An Introduction to Dynamic Meteorology*, 2nd ed. (1979); BRIAN HOSKINS and ROBERT PEARCE (eds.), *Large-Scale Dynamical Processes in the Atmosphere* (1983); S. PANCHEV, *Dynamic Meteorology* (1985; originally published in Bulgarian, 1981); J.V. IRIBARNE and W.L. GODSON, *Atmospheric Thermodynamics*, 2nd ed. (1981); and BARRY SALTZMAN (ed.), *Theory of Climate* (1983). See also HANS A. PANOFSKY and GLENN W. BRIER, *Some Applications of Statistics to Meteorology* (1958), the fundamental text on uses of statistics in atmospheric analysis. (*Micrometeorology*): T.R. OKE, *Boundary Layer Climates* (1978); J.L. MONTEITH (ed.), *Vegetation and the Atmosphere*, 2 vol. (1975–76); and NORMAN J. ROSENBERG, BLAINE L. BLAD, and SHASHI B. VERMA, *Microclimate: The Biological Environment*, 2nd ed. (1983). Turbulence theory is covered in H. TENNEKES and JOHN L. LUMLEY, *A First Course in Turbulence* (1972); and HANS A. PANOFSKY and JOHN A. DUTTON, *Atmospheric Turbulence: Models and Methods for Engineering Applications* (1984). (*Cloud physics and dynamics*): N.H. FLETCHER, *The Physics of Raincloud*s (1962, reprinted 1966); L.T. MATVEEV, *Cloud Dynamics* (1984; originally published in Russian, 1981); and GEORGII HIDY, *Aerosols: An Industrial and Environmental Science* (1984). (*Atmospheric chemistry*): R.D. GOLDBERG (ed.), *Atmospheric Chemistry* (1982); and H.W. GEORGII and W. JAESCHKE, *Chemistry of the Unpolluted and Polluted Troposphere* (1982). (*Air pollution meteorology*): An excellent though now somewhat dated work is JOHN H. SEINFELD, *Air Pollution: Physical and Chemical Fundamentals* (1975). Other useful texts include ARTHUR C. STERN (ed.), *Air Pollution*, vol. 1, *Air Pollutants, Their Transformation and Transport*, 3rd ed. (1976), and vol. 4, *Engineering Control of Air Pollution*, 3rd ed. (1977); and F.T.M. NIEUWSTADT and H. VAN DOP (eds.), *Atmospheric Turbulence and Air Pollution Modelling* (1982, reprinted 1984). (*Synoptic meteorology*): SVERRE PETTERSSEN, *Weather Analysis and Forecasting*, 2nd ed., 2 vol. (1956), is excellent, but much of the material is dated. The best sources are publications of public and military weather services, including RALPH K. ANDERSON *et al.*, *Application of Meteorological Satellite Data in Analysis and Forecasting* (1969, reprinted 1974); UNITED STATES. DEPARTMENT OF THE AIR FORCE. AIR WEATHER SERVICE, *The Use of the Skew T, Log P Diagrams in Analyzing and Forecasting* (1969); and ROBERT C. MILLER, *Notes on Analysis and Severe-Storm Forecasting Procedures of the Air Force Global Weather Central*, rev. ed. (1972). A new forecasting concept is discussed in K.A. BROWNING (ed.), *Nowcasting* (1982). Mesoscale atmospheric systems are discussed in B.W. ATKINSON, *Meso-Scale Atmospheric Circulations* (1981). A summary of current work in applied areas of meteorology, including weather forecasting, is DAVID D. HOUGHTON (ed.), *Handbook of Applied Meteorology* (1985). (*Atmospheric modeling*): The use of computer models to simulate atmospheric features is treated in GEORGE J. HALTINER and ROGER TERRY WILLIAMS, *Numerical Prediction and Dynamic Meteorology*, 2nd ed. (1980); and ROGER A. PIELKE, *Mesoscale Meteorological Modeling* (1984). (*Remote sensing*): RICHARD J. DOVIAK and DUŠAN S. ZRNIC, *Doppler Radar and Weather Observations* (1984); E.C. BARRETT and L.F. CURTIS, *Introduction to Environmental Remote Sensing*, 2nd ed. (1982); E.E. GOSSARD and R.G. STRAUCH, *Radar Observation of Clear Air and Clouds* (1983); and T.D. ALLAN (ed.), *Satellite Microwave Remote Sensing* (1983). (*Weather modification*): Useful summaries are available in W.N. HESS, *Weather and Climate Modification* (1974); and ARNETT S. DENNIS, *Weather Modification by Cloud Seeding* (1980).

(R.A.Pi.)

Earthquakes

An earthquake is a sudden disturbance within the Earth manifested at the surface by a shaking of the ground. This shaking, which accounts for the destructiveness of an earthquake, is caused by the passage of elastic waves through the Earth's rocks. These seismic waves are produced when some form of stored energy, such as elastic strain, chemical energy, or gravitational energy, is released suddenly.

Few natural phenomena can wreak as much havoc as earthquakes. Over the centuries they have been responsible for millions of deaths and an incalculable amount of damage to property. While earthquakes have inspired dread and superstitious awe since ancient times, little was understood about them until the emergence of seismology at the beginning of the 20th century. Seismology, which involves the scientific study of all aspects of earthquakes, has yielded answers to such long-standing questions as why and how earthquakes occur. These matters are discussed in this article, as are the distribution, size, and effects of earthquakes.

For coverage of related topics in the *Macropædia* and *Micropædia*, see the *Propædia*, sections 212, 213, and 231, and the *Index*.

The article is divided into the following sections:

General considerations 655
  Principal types of seismic waves
  Properties of seismic waves
  Seismic instruments and systems
Effects of earthquakes 656
  Primary effects
  Intensity scales
  Tsunamis and seiches
  Some great earthquakes
Causes of earthquakes 658
  Principal mechanisms in nature
  Artificial means of inducing earthquakes
Distribution of earthquakes 660
  Earthquake observatories
  Locating earthquake epicentres
  Geographic concentrations of earthquakes
  Tectonic associations
  Aftershocks, foreshocks, and swarms
  Extraterrestrial seismic phenomena
Size, energy, and frequency of earthquakes 662
  Earthquake magnitude
  Energy and frequency of occurrence
Earthquake prediction 663
  Observation and interpretation of precursory phenomena
  Methods of reducing earthquake hazards
Exploration of the Earth's interior with seismic waves 664
  Seismological methods and earthquake tomography
  Structure of the Earth's interior
  Long-period oscillations of the globe
Bibliography 666

GENERAL CONSIDERATIONS

Principal types of seismic waves. Seismic waves generated by an earthquake source are commonly classified into three main types. The first two, the *P* and *S* waves, are propagated within the Earth, while the third, consisting of Love and Rayleigh waves, is propagated along its surface (Figure 1). The existence of these types of seismic waves was predicted during the 19th century, and modern investigators have found that there is a close correspondence between such theoretical calculations and seismographic measurements of the waves.

P and *S* waves

The *P* (or primary) waves travel through the body of the Earth at the highest speeds. They are longitudinal waves that can be transmitted by both solid and liquid materials in the Earth's interior. With *P* waves, the particles of the medium vibrate in a manner similar to sound waves, and the transmitting rocks are alternately compressed and expanded. The other type of body wave, the *S* (or secondary) wave, travels only through solid material within the Earth. With *S* waves, the particle motion is transverse to the direction of travel and involves the shearing of the transmitting rock.

Because of their greater speed, the *P* waves are the first to reach any point on the Earth's surface. The first *P*-wave onset starts from the spot where an earthquake originates. This point, usually at some depth within the Earth, is called the focus, or hypocentre. The point immediately above the focus at the surface is known as the epicentre.

Love and Rayleigh waves

Love and Rayleigh waves are guided by the free surface of the Earth. They follow along after the *P* and *S* waves have passed through the body of the planet. Both Love and Rayleigh waves involve horizontal particle motion, but only the latter type has vertical ground displacements. As Love and Rayleigh waves travel, they disperse into long wave trains, and at substantial distances from the source they cause much of the shaking felt during earthquakes.

Properties of seismic waves. At all distances from the focus, the mechanical properties of the rocks, such as incompressibility, rigidity, and density, play a role in the speed with which the waves travel and the shape and duration of the wave trains. The layering of the rocks and the physical properties of surface soil also affect these characteristics of the waves. In most cases, elastic behaviour occurs in earthquakes, but the shaking of surface soils from the incident seismic waves sometimes results in nonelastic behaviour, including slumping (*i.e.*, the downward and outward movement of unconsolidated material) and the liquefaction of sandy soil.

When a seismic wave encounters an interface or boundary that separates rocks of different elastic properties, it undergoes reflection and refraction. There is a special complication if a conversion between the wave types occurs at such a boundary: either an incident *P* or *S* wave can yield in general reflected *P* and *S* waves and refracted *P* and *S* waves. Boundaries between structural layers also give rise to diffracted and scattered waves. These additional waves are in part responsible for the complications observed in ground motion during earthquakes. Modern research is concerned with computing, from the theory of waves in complex structures, synthetic records of ground motion that are realistic in comparison with observed ground shaking.

The frequency range of seismic waves is large. Seismic waves may have frequencies from as high as the audible range (greater than 20 hertz [Hz]) to as low as the free oscillations of the whole Earth, with gravest period being 54 minutes (*i.e.*, the Earth vibrates in various modes, and the mode with the lowest pitch takes 54 minutes to complete a single vibration; see below *Long-period oscillations of the globe*). Attenuation of the waves in rock imposes high-frequency limits, and in small to moderate earthquakes measured surface waves have frequencies extending from about one to 0.005 Hz.

The amplitude range of seismic waves is also great in most earthquakes. The displacements of the ground extend from 10^{-10} to 10^{-1} metres. In the greatest earthquakes, the ground amplitude of the predominant *P* waves may be several centimetres at periods of two to five seconds. Very close to the seismic sources of great earthquakes, investigators have measured large wave amplitudes with

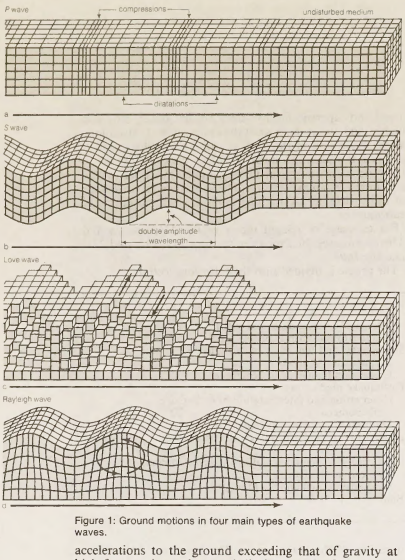


Figure 1: Ground motions in four main types of earthquake waves.

accelerations to the ground exceeding that of gravity at high frequencies and ground displacements of one metre at low frequencies.

Seismic instruments and systems. Ground motion in earthquakes and microseisms (small, often long-continuing oscillations of the ground that do not originate in earthquakes) are both recorded by seismographs. Most of these instruments are of the pendulum type. Still in use today are mechanical seismographs that have a pendulum of large mass (up to several tons) and that produce seismograms by scratching a line on smoked paper on a rotating drum. In more advanced instruments, seismograms are recorded by means of a ray of light from the mirror of a galvanometer through which passes an electric current generated by electromagnetic induction when the pendulum of the seismograph moves. Technological developments, notably in electronics, have given rise to high-precision pendulum seismometers and sensors of ground motion. In these instruments, the electric voltages produced by motions of the pendulum or the equivalent are passed through electronic circuitry to amplify the ground motion for more exact readings.

Major types of seismographs

Generally speaking, seismographs are divided into three types: short period; long (or intermediate) period; and ultra-long period, or broad-band, instruments. Short-period instruments are used to record *P*- and *S*-body waves with high magnification of the ground motion. For this purpose, the seismograph response is shaped to peak at a period of about one second or less. The long- or intermediate-period instruments of the type used by the World-Wide Standard Seismographic Network (WWSSN; see below) have a response maximum at about 20 seconds. Again, in order to provide as much flexibility as possible for research work, the trend has been toward the operation of very-broad-band seismographs, often with digital representation of the signals. This is usually accomplished with very-long-period pendulums and electronic amplifiers that pass signals in the 0.005 to 50 Hz band.

When seismic waves close to their source are to be recorded, special design criteria are needed. Instrument sensitivity must ensure that the largest ground movements remain on scale. For most seismological and engineering purposes the wave frequency is high, and so the pendulum or its equivalent can be small. For comparison, displacement meters need a long free period and pendulum with consequent instability. Accelerometers that measure the rate at which the ground velocity is changing have an advantage for strong-motion recording, because they allow integration to be carried out to estimate ground velocity and displacement. The ground accelerations to be registered range up to twice gravity ($2g$). Recording such accelerations can be easily accomplished with short torsion suspensions or force-balance mass-spring systems.

Because many strong-motion instruments need to be placed at unattended sites in ordinary buildings for periods of months or years before a strong earthquake occurs, they usually record only when a trigger mechanism is actuated with the onset of motion. Solid-state memories are now used, particularly with digital recording instruments, making it possible to preserve the first few seconds before the trigger starts the permanent recording. In the past, recordings were usually made on film strips for up to a few minutes' duration. In present-day equipment, digitized signals are stored directly on magnetic cassette tape or on a memory chip. In most cases absolute timing has not been provided on strong-motion records but only accurate relative time marks; the present trend, however, is to provide Universal Time (the local mean time of the prime meridian) by means of special radio receivers or small crystal clocks.

The prediction of strong ground motion and response of engineered structures in earthquakes depends critically on measurements of the spatial variability of earthquake intensities near the seismic wave source. In an effort to secure such measurements, special arrays of strong-motion seismographs are being installed in areas of high seismicity around the world. Large-aperture seismic arrays (linear dimension on the order of one to 10 kilometres) of strong-motion accelerometers can now be used to improve estimations of speed, direction of propagation, and type of seismic wave components. Like an array of radio telescopes, a seismic array allows wave correlations for consecutive time and frequency intervals so that variations in shaking over small-to-moderate distances can be measured.

Seismic arrays

Finally, because 70 percent of the Earth's surface is covered by water, there is a need for ocean-bottom seismometers to augment the global land-based system of recording stations. Research is under way to determine the feasibility of extensive long-term recording by instruments on the seafloor. Japan already has a semipermanent seismograph system of this type. The system was placed on the seafloor off the Pacific coast of central Honshu in 1978 by means of a cable.

Because of the mechanical difficulties of maintaining permanent ocean-bottom instrumentation, different systems have been considered. These include instruments that are placed in an ocean-bottom package; signals from the instruments are either transmitted to the ocean surface for retransmission by auxiliary apparatus or transmitted via cable to a shore-based station. Another system is designed to release automatically its recording component, allowing it to float to the surface for later recovery.

The use of ocean-bottom seismographs should yield much improved global coverage of seismic waves and provide important information on the seismicity of oceanic regions. Ocean-bottom seismographs will enable investigators to determine the details of the crustal structure of the seafloor and, because of the relative thinness of the oceanic crust, should make it possible for them to collect clear seismic information about the upper mantle. Such systems are also expected to provide new data on focal mechanism, on the origin and propagation of microseisms, and on the nature of ocean-continent margins.

EFFECTS OF EARTHQUAKES

Primary effects. Earthquakes have varied effects, including changes in geologic features, damage to man-made structures, and impact on human and animal life.

Geomorphological changes caused by earthquakes

Geomorphological changes are often caused by an earthquake: *e.g.*, movements—either vertical or horizontal—along geological fault traces; the raising, lowering, and tilting of the ground surface with related effects on the flow of groundwater; liquefaction of sandy ground; landslides; and mudflows. The investigation of topographical changes is aided by geodetic measurements, which are made systematically in a number of countries seriously affected by earthquakes.

Earthquakes can do significant damage to buildings, bridges, pipelines, railways, embankments, and other manmade structures. The type and extent of damage inflicted are related to the strength of the ground motions and to the behaviour of the foundation soils.

In the most intensely damaged region, called the meizoseismal area, the effects of a severe earthquake are usually complicated and depend on the topography and the nature of the surface materials; they are often severer on soft alluvium and unconsolidated sediments than on hard rock. At distances of more than 100 kilometres (62 miles) from the source, the main damage is caused by surface waves. In mines there is frequently little damage below depths of a few hundred metres even though the surface immediately above is considerably affected.

Further effects of interest are the occurrence of earthquake sounds and lights. The sounds are generally low-pitched and have been likened to the noise of an underground train passing through a station. The occurrence of such sounds implies the existence of significant short periods in the *P* waves in the ground (a wave period is the length of time between the arrival of successive crests in a wave train). Occasionally luminous flashes, streamers, and balls are seen in the night sky during earthquakes. These lights have been attributed to electric induction in the air along the earthquake source.

Intensity scales. The level of violence of seismic shaking varies considerably over the affected area. This intensity is not capable of simple quantitative definition and, particularly before seismographs capable of accurate measurement of ground motion were developed, the shaking was estimated by reference to intensity scales that describe the effects in qualitative terms. Subsequently, the divisions in these scales have been associated with accelerations of the local ground shaking. Intensity depends, however, in a complicated way not only on ground accelerations but also on the periods and other features of seismic waves, the distance of the point from the source, and the local geological structure. Furthermore, it is distinct from magnitude, which is a measure of earthquake size specified by a seismograph reading (see below *Earthquake magnitude*).

A number of different intensity scales have been set up during the past century and applied to both current and ancient destructive earthquakes. For many years the most widely used was the 10-point scale devised by Michele Stefano de Rossi and François-Alphonse Forel in 1878.

Mercalli scale

The scale now generally employed in North America is the Mercalli scale, as modified by Harry O. Wood and Frank Neumann in 1931, in which intensity is considered to be more uniformly graded. An abridged form of the modified Mercalli scale is provided below. Alternative scales have been developed in both Japan and Europe for local conditions. The European (MSK) scale of 12 grades is similar to the abridged version of the Mercalli.

Modified Mercalli Scale of Felt Intensity (1931; Abridged)

I. Not felt. Marginal and long-period effects of large earthquakes.

II. Felt by persons at rest, on upper floors, or otherwise favourably placed to sense tremors.

III. Felt indoors. Hanging objects swing. Vibrations like passing of light trucks. Duration can be estimated.

IV. Vibration like passing of heavy trucks (or sensation of a jolt like a heavy ball striking the walls). Standing motorcars rock. Windows, dishes, doors rattle. Glasses clink. Crockery clashes. In the upper range of IV, wooden walls and frames creak.

V. Felt outdoors; direction may be estimated. Sleepers wakened. Liquids disturbed, some spilled. Small objects displaced or upset. Doors swing, open, close. Pendulum clocks stop, start, change rate.

VI. Felt by all; many frightened and run outdoors. Persons walk unsteadily. Pictures fall off walls. Furniture moved or overturned. Weak plaster and masonry cracked. Small bells ring (church, school). Trees, bushes shaken.

VII. Difficult to stand. Noticed by drivers of motorcars. Hanging objects quiver. Furniture broken. Damage to weak masonry. Weak chimneys broken at roof line. Fall of plaster, loose bricks, stones, tiles, cornices. Waves on ponds; water turbid with mud. Small slides and caving along sand or gravel banks. Large bells ring. Concrete irrigation ditches damaged.

VIII. Steering of motorcars affected. Damage to masonry; partial collapse. Some damage to reinforced masonry; none to reinforced mansonry designed to resist lateral forces. Fall of stucco and some masonry walls. Twisting, fall of chimneys, factory stacks, monuments, towers, elevated tanks. Frame houses moved on foundations if not bolted down; loose panel walls thrown out. Decayed piling broken off. Branches broken from trees. Changes in flow or temperature of springs and wells. Cracks in wet ground and on steep slopes.

IX. General panic. Weak masonry destroyed; ordinary masonry heavily damaged, sometimes with complete collapse; reinforced masonry seriously damaged. Serious damage to reservoirs. Underground pipes broken. Conspicuous cracks in ground. In alluvial areas, sand and mud ejected, earthquake fountains, sand craters.

X. Most masonry and frame structures destroyed with their foundations. Some well-built wooden structures and bridges destroyed. Serious damage to dams, dikes, embankments. Large landslides. Water thrown on banks of canals, rivers, lakes, etc. Sand and mud shifted horizontally on beaches and flat land. Railway rails bent slightly.

XI. Rails bent greatly. Underground pipelines completely out of service.

XII. Damage nearly total. Large rock masses displaced. Lines of sight and level distorted. Objects thrown into air.

With the use of an intensity scale, it is possible to summarize the macroseismic data for an earthquake by constructing isoseismal curves, which are the loci of points that demarcate areas of equal intensity. If there were complete symmetry about the vertical through the earthquake's focus, the isoseismals would be circles with the epicentre as centre. However, because of the many unsymmetrical factors influencing the intensity, the curves are often far from circular.

Area of highest intensity

The most probable position of the epicentre based on macroseismic data will be at a point inside the area of highest intensity. In some cases, it is verified by instrumental data that the epicentre is satisfactorily determined in this way, but not infrequently the true epicentre lies outside the area of greatest intensity.

Tsunamis and seiches. *Tsunamis.* Very long water waves in oceans or seas, tsunamis (or seismic sea waves), sweep inshore following certain earthquakes. They sometimes reach great heights and may be extremely destructive. The immediate cause of a tsunami is a disturbance in an adjacent seabed sufficient to cause the sudden raising or lowering of a large body of water. This disturbance may be centred in the focal region of an earthquake or it may be a submarine landslide arising from an earthquake. Following the initial disturbance to the sea surface, water waves spread out in all directions. Their speed of travel in deep water is given by $(gh)^{\frac{1}{2}}$, where h is the sea depth and g is the acceleration of gravity. This speed may be considerable; *e.g.*, 100 metres per second (224 miles per hour) when h is 1,000 metres (3,280 feet). The amplitude at the surface does not exceed a few metres in deep water, but the principal wavelength may be on the order of hundreds of kilometres; correspondingly, the principal wave period may be on the order of tens of minutes. Because of these features, the waves are not noticed by ships far out at sea.

Speed of travel

When tsunamis approach shallow water, the wave amplitude increases. The waves may occasionally reach a height of 20 to 30 metres in U- and V-shaped harbours and inlets. They sometimes do a great deal of damage in low-lying ground around such inlets. Frequently the wave front in the inlet is nearly vertical, as, for example, in a tidal bore, and the speed of onrush may be on the order of 10 metres per second. In some cases there are several great waves separated by intervals of several minutes or more. The first of these waves is often preceded by an extraordinary recession of water from the shore, which may commence several minutes or even half an hour beforehand.

Organizations, notably in Japan, Siberia, Alaska, and Hawaii, have been set up to provide tsunami warnings. A key development is the Seismic Sea Wave Warning System (SSWWS), an internationally supported system designed to reduce loss of life in the Pacific Ocean. Centred in Honolulu, it issues alerts based on reports of earthquakes from circum-Pacific seismographic stations.

Seiches. These are rhythmic motions of water in nearly landlocked bays or lakes that are sometimes induced by earthquakes and by tsunamis (in the case of the former). Oscillations of this sort may last for hours or even for a day or two.

The great Lisbon earthquake of 1755 caused the waters of canals and lakes in areas as far away as Scotland and Sweden to go into observable oscillations. Seiche surges in Texas in the southwestern United States commenced between 30 and 40 minutes after the 1964 Alaska earthquake and were produced by seismic surface waves passing through the area.

Of course, *P* waves from an earthquake may pass through the sea following refraction through the seafloor. The speed of these waves is about 1.5 kilometres per second, the speed of sound in water. If such waves meet a ship with sufficient intensity, they give the impression that the ship has struck a submerged object. This phenomenon is called a seaquake.

Some great earthquakes. About 50,000 earthquakes large enough to be felt or noticed without the aid of instruments occur annually over the entire Earth. Of these, approximately 100 are of sufficient size to produce substantial damage if their centres are near areas of habitation. Very great earthquakes occur at an average rate of about one per year. Among the great earthquakes of the past are those of Lisbon in 1755; New Madrid, Mo., U.S., in December 1811 and January and February 1812; San Francisco in 1906; Tokyo–Yokohama in 1923; the coast of Chile in 1960; south-central Alaska in 1964; T'ang-shan, China, in 1976; and Mexico in 1985. Their devastating effects are briefly described below.

Lisbon. On Nov. 1, 1755, Lisbon was heavily damaged by a great earthquake that occurred at 9:40 AM. The source was situated some distance off the coast. The violent shaking demolished large public buildings and about 12,000 dwellings. As November 1 was All Saint's Day, a large part of the population was attending religious services; most of the churches were destroyed, resulting in many casualties. The total number of persons killed in Lisbon alone was estimated to be as high as 60,000, including those who perished by drowning and in the fire that burned for about six days following the shock. Damage was reported in Algiers, 1,100 kilometres to the east. The earthquake generated a tsunami that produced waves about six metres high at Lisbon and 20 metres high at Cádiz, Spain. The waves traveled on to Martinique, a distance of 6,100 kilometres in 10 hours, and there rose to a height of four metres.

New Madrid. Three large earthquakes occurred near New Madrid in southern Missouri on Dec. 16, 1811, and Jan. 23 and Feb. 7, 1812. There were numerous aftershocks, of which 1,874 were large enough to be felt in Louisville, Ky., some 300 kilometres away. The principal shock produced waves of sufficient amplitude to shake down chimneys in Cincinnati, Ohio, about 600 kilometres away. The waves were felt as far as Canada in the north and the Gulf Coast in the south. The area of greatest shaking was about 100,000 square kilometres, considerably greater than the area involved in the San Francisco earthquake in 1906. It has been discovered that in continental earthquakes such as the Missouri shocks, the area of strong shaking can be abnormally large compared with that in shocks along the Pacific coast of the United States. In one region 240 kilometres long by 60 kilometres wide, the ground sank from one to three metres and was covered by inflowing river water. Sand liquefaction effects were widespread. In certain locations, forests were overthrown or ruined by the loss of soil shaken from the roots of the trees.

San Francisco. On April 18, 1906, at about 5:12 AM, the San Andreas Fault slipped over a segment about 430 kilometres long, extending from San Juan Bautista in San Benito County to the upper Mattole River in Humboldt County and from there perhaps out under the sea to an unknown distance. The shaking was felt from Los Angeles in the south to Coos Bay, Ore., in the north. Damage was severe in San Francisco and in other towns situated near the fault—*e.g.*, San Jose, Salinas, and Santa Rosa (30 kilometres from the fault). Approximately 700 people were killed. In San Francisco the earthquake started a fire, which destroyed the central business district.

Tokyo–Yokohama. A great earthquake struck the Tokyo–Yokohama metropolitan area near noon on Sept. 1, 1923. The death toll from this shock was estimated at more than 140,000. Fifty-four percent of the brick buildings and 10 percent of the reinforced concrete structures collapsed. Many hundreds of thousands of houses were either shaken down or burned. The shock started a tsunami that reached a height of 12 metres at Atami on Sagami-nada (Sagami Gulf), where it destroyed 155 houses and killed 60 persons.

Chile. The source of this earthquake in 1960 extended over a distance of about 1,100 kilometres along the southern Chilean coast. Casualties included about 5,700 killed and 3,000 injured, and property damage amounted to many millions of dollars. Seismic sea waves excited by the earthquake caused death and destruction in Hawaii, Japan, and the Pacific coast of the United States.

Alaska. On March 27, 1964, a great earthquake with a Richter magnitude 8.3–8.5 (see below) occurred in south central Alaska. It released at least twice as much energy as the 1906 San Francisco earthquake and was felt on land over an area of almost 1,300,000 square kilometres. The death toll was only 131 because of the low density of the state's population, but property damage was very high. The earthquake tilted an area of at least 120,000 square kilometres. Landmasses were thrust up locally as high as 25 metres to the east of a line extending northeastward from Kodiak Island through the western part of Prince William Sound. To the west, land sank as much as 2.5 metres. Extensive damage in coastal areas resulted from submarine landslides and tsunamis. Tsunami damage occurred as far away as Crescent City, Calif. The occurrence of tens of thousands of aftershocks indicates that the region of faulting extended about 1,000 kilometres.

T'ang-shan. The coal-mining and industrial city of T'ang-shan, located about 110 kilometres east of Peking, was almost razed in the tragic earthquake of July 28, 1976. The death toll exceeded 240,000 persons, and probably another 500,000 were injured. Most persons were killed from the collapse of unreinforced masonry homes, where they were asleep.

Mexico. The main shock occurred at 7:18 AM on Sept. 19, 1985. The cause was a fault slip along the Benioff zone (a band of intermediate- and deep-earthquake foci along a planar dipping zone) under the Pacific coast of Mexico. Although 400 kilometres from the epicentre, Mexico City suffered major building damage and more than 10,000 of its inhabitants were reported killed. The highest intensity was in the central city, which is founded on a former lake bed. The ground motion there measured five times that in the outlying districts, which have different soil foundations.

CAUSES OF EARTHQUAKES

Principal mechanisms in nature. Earthquakes are caused by the sudden release of energy within some limited region of the rocks of the Earth. The form of energy involved is produced by elastic strain, gravitational potential, chemical reactions, or motion of bodies. Of these, the release of elastic strain energy is the most important, since this form of energy is the only kind that is stored in sufficient quantity in the Earth to produce major earthquakes. Earthquakes associated with this type of energy release are called tectonic earthquakes.

Measurements of triangulation lines across the San Andreas Fault before and after its rupture in the 1906 San Francisco earthquake led to the so-called elastic rebound theory for tectonic earthquakes. As formulated by the American geologist Harry Fielding Reid, the theory ex-

Elastic rebound theory

plains that a tectonic earthquake occurs when stresses in rock masses have accumulated to a point where they exceed the strength of the rocks, leading to rapid fracture. These rock fractures usually tend in the same direction and may extend over many kilometres along the zone of weakness. In the 1906 earthquake the San Andreas Fault slipped for 430 kilometres, with a maximum horizontal fault offset of about six metres.

Another type of earthquake, that associated with volcanic activity, is called a volcanic earthquake. Yet, it is likely that even here the energy released may be the result of a relatively sudden slip of rock masses and the consequent release of elastic strain energy. The energy, however, may in part be of hydrodynamic origin due to the motion of magma in reservoirs beneath the volcano or to the release of gas under pressure.

The elastic rebound theory of an earthquake source envisages the flinging of rock masses in opposite directions on each side of the rupturing fault as the fault rupture progresses along the fault. In the rupture, the rock masses spring back to a position where the elastic strain is less. This movement at any point may not take place at once but rather in irregular steps. These sudden stoppings and startings give rise to the vibrations that propagate as seismic waves. The irregular properties of fault rupture are now included in the modeling of earthquake sources, both physically and mathematically. Roughnesses along the fault are referred to as asperities, and places where the rupture slows or stops are said to be fault barriers. Fault rupture starts at the earthquake focus and propagates unilaterally or bilaterally over the fault plane until stopped or slowed at a barrier. The result is a redistribution of elastic strain, which may or may not break the barrier. Sometimes the fault rupture is reinstated on the far side of the barrier; sometimes the stresses in the rocks eventually produce a breakage, and the rupture continues.

Earthquakes have different properties depending on the type of fault slip that causes them. The geological interpretation of a fault is given in terms of standard geometries (Figure 2). The usual fault model has a strike (direction from north of the horizontal line in the fault plane) and a dip (angle between direction of steepest slope and horizontal). The hanging wall lies over the footwall, the lower wall of an inclined fault.

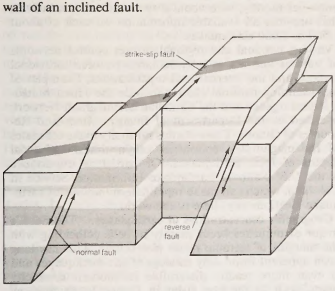


Figure 2: Types of faulting.

Types of fault movement

Relative offsets parallel to the strike produce strike-slip faulting while those parallel to the dip generate dip-slip faulting. Strike-slip faults are right or left lateral, depending on whether the block on the opposite side of the fault from the observer moves to his right or left. Dip-slip faults are normal if the hanging-wall block moves downward relative to the footwall block; the opposite motion produces reverse or thrust faulting. A mixed offset results in oblique-slip faulting, which is measured either by the plunge or by the slip angle.

Observed faults are assumed to be the seat of one or more past earthquakes, though movements along faults are often slow, and most geologically ancient faults are now aseismic (*i.e.*, cause no earthquakes). The actual faulting in an earthquake may be complex, and it is often not clear whether in a particular earthquake the total energy issues from a single fault plane.

Observed geological faults sometimes show overall relative displacements on the order of hundreds of kilometres, whereas the amplitudes of seismic waves reach only several centimetres. In the 1976 T'ang-shan earthquake, for example, a surface strike-slip of about one metre was observed along the causative fault.

Polarity method

An important research technique is to infer the character of faulting in an earthquake from observed distributions of the directions of the first onsets in waves arriving at the Earth's surface. Onsets have been called compressional or dilatational according to whether the direction is away from or toward the focus, respectively. A polarity pattern becomes recognizable when the directions of the *P*-wave onsets are plotted on a map: there are broad areas in which the first onsets are predominantly compressions, separated from predominantly dilatational areas by nodal curves near which the *P*-wave amplitudes are abnormally small.

In 1926 the American geophysicist Perry E. Byerly used patterns of *P* onsets over the entire globe to infer the orientation of the fault plane in a large earthquake. The polarity method yields two *P*-nodal curves at the Earth's surface. For a homogeneous Earth, one curve is in the plane containing the assumed fault, and the other is in the plane (called the auxiliary plane) that passes through the focus and is perpendicular to the forces of the plane. For the actual Earth, the nodal curves are displaced from these locations because of the curvature of the wave paths between focus and surface, but knowledge of Earth structure enables allowance to be made for this. Given an adequately well-determined pattern of first *P*-wave movements, it is possible to locate two planes, one of which is the plane containing the fault.

Artificial means of inducing earthquakes. Earthquakes are sometimes caused by human activities. Such activities include the injection of fluids into deep wells, the detonation of large underground nuclear explosions, the excavation of mines, and the filling of large reservoirs. In the case of deep mining, the removal of rock produces changes in the strain around the tunnels. Slip on preexisting faults or outward shattering of rock into the cavities may occur. In all other situations, the induction mechanism is thought to involve elastic strain release, as in the case of tectonic earthquakes. Here, earthquakes are triggered by small changes in the local strain field that produce rock fracture or fault slip. Local changes in strain around large underground explosions have been known to produce slip on already strained faults in the vicinity.

Reservoir induction. Of the various activities cited above, the filling of large reservoirs is among the most important. More than 20 cases have been documented in which local seismicity has increased following the impounding of water behind high dams. Other claims cannot be substantiated because the necessary observations that allow comparison of earthquake occurrence before and after filling do not exist. Reservoir-induction effects are most marked for reservoirs exceeding 100 metres in depth and one cubic kilometre in volume. Three cases where such effects have very probably been involved are the Hoover Dam in the United States, the Aswān High Dam in Egypt, and the Kariba Dam on the border between Zimbabwe and Zambia. The most generally accepted explanation for the cause of the earthquake occurrence in such cases is that rocks near the reservoir are already strained from the regional tectonic forces to a point where nearby faults are almost ready to slip. Water in the reservoir adds a pressure perturbation that triggers the fault rupture. The pressure effect perhaps is enhanced by the fact that the rocks along the fault have lower strength due to increased rocks along the fault have lower strength due to increased water-pore pressure. These factors notwithstanding, it has been determined that most large reservoirs do not produce earthquakes.

The specific earthquake mechanisms associated with reservoir induction have been established in a few cases. For the main shock at the Koyna Dam and Reservoir in India, the evidence favours strike-slip motion, and at Hsinfeng-chiang Dam in China, the principal shock can also be

attributed to the strike-slip mechanism. At both the Kremasta Dam in Greece and the Kariba Dam in Zimbabwe–Zambia, the mechanism was dip-slip on normal faults. By contrast, thrust mechanisms have been determined for earthquakes at the lake behind Nurek Dam in the Soviet Union. More than 1,800 earthquakes occurred during the first nine years after water was impounded in this 317-metre deep reservoir, a rate amounting to four times the average number of shocks in the region prior to filling.

Seismology and nuclear explosions. In 1958 representatives from several countries, including the United States and the Soviet Union, met to discuss the technical basis for a nuclear test ban treaty. Among the matters considered was the feasibility of developing effective means with which to detect underground nuclear explosions and to distinguish them seismically from earthquakes. Since that conference, much attention has been devoted to seismological research, leading to major advances in seismic signal detection and analysis in terms of both instrumentation and methodology.

Stimulus for seismological research

Recent seismological work on test ban treaty verification has involved using high-resolution seismographs, estimating the yield of explosions, studying wave attenuation in the Earth, determining wave amplitude and frequency spectra discriminants, and applying seismic arrays (see above). The findings of such research have shown that underground nuclear explosions, compared with natural earthquakes, usually generate larger amplitude *P* waves relative to the surface waves. The extension of seismic explosion research (and the experimental controls that go with it) to seismological problems has yielded useful information on seismic wave propagation in general and on the Earth's structure.

DISTRIBUTION OF EARTHQUAKES

Earthquake observatories. During the late 1950s there existed worldwide only about 700 seismographic stations equipped with seismographs of various types and frequency responses. Few instruments were calibrated, so that actual ground motions could not be measured and timing errors of several seconds were common. The WWSSN, the aforementioned worldwide standardized seismographic network, was established to help remedy this situation. Each station of the WWSSN has six seismographs—three short-period and three long-period seismographs. Timing and accuracy are maintained by crystal clocks, and a calibration pulse is placed daily on each record. By 1967 the WWSSN consisted of about 120 stations distributed over 60 countries. Other countries, such as Canada, which did not participate directly in the WWSSN, upgraded their own stations in order to make them compatible with the standardized network. The resulting data provided the basis for significant advances in research on earthquake mechanisms, global tectonics, and the structure of the Earth's interior.

By the 1980s a further upgrading of permanent seismographic stations had begun with the installation of digital equipment. Among the global networks of digital seismographic stations now in operation are the seismic research observatories in boreholes 100 metres deep; modified high-gain, long-period (surface) observatories; and digital worldwide standardized seismographic network (DWWSSN) stations. In addition, a number of gravimeters capable of digital recording and response to very long wavelengths have been installed throughout the world as part of the International Deployment of Accelerographs (IDA) network. The main aim is to equip global observatories with seismographs that can record seismic waves over a broad band of frequencies.

Networks of digital seismographic stations

Locating earthquake epicentres. At some observatories it is customary to make provisional estimates of the epicentres of the more important earthquakes. These estimates provide preliminary information locally about particular earthquakes and serve as first approximations for the calculations subsequently made by large coordinating centres. (Distances between stations and an earthquake epicentre are calculated in terms of the angle subtended at the Earth's centre; Figure 5.)

In the case of a single observatory, an earthquake's epicentre can often be estimated from the readings of three perpendicular component seismograms. For example, for a shallow earthquake the epicentral distance, if less than 105°, is indicated by the interval between the arrival times of *P* and *S* waves; the azimuth and angle of emergence are indicated by a comparison of the sizes and directions of the first movements shown in the seismograms and by the relative sizes of later waves, particularly surface waves. It should be noted, however, that in certain regions the first wave movement at a station arrives from a direction differing from the azimuth toward the epicentre. The explanation is usually in terms of strong variations in geological structures.

When data from more than one observatory are available, an earthquake's epicentre may be estimated from the epicentral distances indicated by the times of travel of the *P* and *S* waves from source to recorder. Nowadays, in many seismically active regions, networks of seismographs with telemetry transmission and centralized timing and recording are common. Whether analog or digital recording is used, such integrated systems greatly simplify observatory work: multichannel signal displays make identification and timing of phase onsets easier and more reliable.

Moreover, modern on-line microprocessors can be programmed to pick automatically, with some degree of confidence, the onset of a significant common phase, such as *P*, by correlation of waveforms from parallel network channels. With the aid of specially designed computer programs, seismologists can then locate distant earthquakes to within about 10 kilometres and the epicentre of a local earthquake to within just a few kilometres.

Catalogs of felt earthquakes and earthquake observations have appeared intermittently for many centuries. The earliest known list of instrumentally recorded earthquakes with computed times of origin and epicentres is that for the period 1899–1903. In subsequent years, cataloging of earthquakes has become increasingly more uniform and complete. Especially valuable is the service provided by the International Seismological Centre (ISC) at Newbury, Eng. Each month it receives about 80,000 readings from about 1,200 stations worldwide and preliminary estimates of the locations of approximately 1,600 earthquakes from national and regional agencies and observatories. The ISC publishes monthly, with about a two-year delay, a bulletin that provides all available information on each of about 1,500 to 2,000 earthquakes.

Cataloging of earthquakes

Various national and regional centres control networks of stations and act as intermediaries between individual stations and the international organizations. Examples of long-standing national centres include the Japan Meteorological Agency and the Canadian Seismograph Network operated by the Department of Energy, Mines and Resources of Ottawa. These centres normally make estimates of the magnitudes, epicentres, origin times, and focal depths of local earthquakes. Of particular importance is the U.S. National Earthquake Information Service in Colorado, which can make rapid determinations of earthquake locations anywhere in the world.

Geographic concentrations of earthquakes. The Earth's major earthquakes occur mainly in belts coinciding with the margins of tectonic plates (see below). This has long been apparent from early catalogs of felt earthquakes and is even more readily discernible in modern seismicity maps, such as the one given in Figure 3, which show instrumentally determined epicentres.

One major earthquake belt passes around the Pacific Ocean and affects coastlines bordering on it, as, for example, those of New Zealand, New Guinea, Japan, the Aleutian Islands, Alaska, and the western regions of North and South America. It is estimated that 80 percent of the energy presently released in earthquakes comes from those whose epicentres are in this belt. The seismic activity is by no means uniform throughout the belt, and there are a number of branches at various points.

Major earthquake belts

A second belt passes through the Mediterranean region eastward through Asia and joins the first belt in the East Indies. The energy released in earthquakes from this belt is about 15 percent of the world total. There also are striking connected belts of seismic activity, mainly along



Figure 3: Distribution of earthquake epicentres (black areas), 1963–77.

mid-oceanic ridges—including those in the Arctic Ocean, the Atlantic Ocean, and the western Indian Ocean—and along the rift valleys of East Africa.

Most other parts of the world experience at least occasional shallow earthquakes—those that originate within 60 kilometres of the Earth's outer surface. The great majority of earthquakes are shallow. It should be noted that the geographic distribution of smaller earthquakes is less precisely determined, partly because the availability of relevant data is dependent on the geographical distribution of observatories.

A distinction is made between "intermediate" focal depths ranging from about 60 to 300 kilometres and greater focal depths. Of the total energy released in earthquakes, 12 percent comes from intermediate earthquakes and 3 percent from deeper ones. The frequency of occurrence falls off rapidly with increasing focal depth in the intermediate range, while below this the distribution in depth is fairly uniform until the greatest focal depths are approached.

Concentrations of deep-focus earthquakes

Deep-focus earthquakes commonly occur in patterns called Benioff zones that dip into the Earth. Dip angles average about 45°, with some shallower and others nearly vertical. Benioff zones are found under tectonically active island arcs, such as Japan, Vanuatu (formerly the New Hebrides), the Kingdom of Tonga (islands), and Alaska, and they are normally but not always (*e.g.*, Romania and the Hindu Kush mountain system) associated with deep ocean trenches, such as those along the South American Andes. In most Benioff zones intermediate- and deep-earthquake foci lie in a narrow layer, although recent precise hypocentral locations in Japan and elsewhere show two distinct parallel bands of foci 20 kilometres apart. Careful estimation gives about 680 kilometres for the deepest depths globally.

Tectonic associations. There is a clear correspondence between the geographical distribution of volcanoes and major earthquakes, particularly in the circum-Pacific earthquake belts and along mid-oceanic ridges. Volcanic vents, however, are generally at a distance of some hundreds of kilometres from the majority of the epicentres of major shallow earthquakes, and many earthquake sources occur nowhere near active volcanoes. Earthquakes of intermediate focal depth frequently occur directly below structures marked by volcanic vents, but there is probably no immediate causal connection between these earthquakes and the volcanic activity, both most likely resulting from the same tectonic processes.

Seismicity patterns had no strong global theoretical explanation until a dynamical model called plate tectonics was developed during the late 1960s. This theory holds that the Earth's upper shell, or lithosphere, consists of nearly a dozen large, quasi-stable slabs called plates. The thickness of each of these plates extends to a depth of roughly 80 kilometres. The plates move horizontally, relative to neighbouring plates, on a layer of softer rock. The rate of movement ranges from one to 10 centimetres per year over a shell of lesser strength called the asthenosphere. At the plate edges where there is contact with adjoining plates, boundary tectonic forces operate on the rocks, causing physical and chemical changes in them. New lithosphere is created at mid-oceanic ridges by the upwelling and cooling of magma from the Earth's mantle. The horizontally moving plates are believed to be absorbed at the ocean trenches, where a subduction process carries the lithosphere downward along the Benioff zones into the Earth's interior. The total amount of lithospheric material destroyed at these subduction zones equals that generated at the ridges.

Seismological evidence (*e.g.*, location of major earthquake belts, as shown in Figure 3) is broadly in agreement with this kinematic model. Earthquake sources are concentrated along the midoceanic ridges, which correspond to divergent plate boundaries. At the subduction zones, which are associated with convergent plate boundaries, intermediate- and deep-focus earthquakes in the Benioff zone mark the location of the upper part of a dipping plate. The focal mechanisms indicate that the stresses are aligned with the dip of the lithosphere underneath the adjacent continent or island arc.

High seismicity along plate boundaries

Some earthquakes associated with mid-oceanic ridges are confined to strike-slip faults that offset the ridge crests. The majority of the earthquakes occurring along such horizontal shear faults are characterized by slip motions. Also consistent with the plate tectonics theory is the high seismicity encountered along the edges of plates that slide past each other. Examples of plate boundaries of this kind, which are sometimes called fracture zones, include the San Andreas Fault in California and the North Anatolian fault system in Turkey. Such plate boundaries are the site of interplate earthquakes of shallow focus.

One other point that correlates with the plate theory is the low seismicity within plates. Small to large earthquakes do occur in limited regions well within the boundaries of plates; however, such interplate seismic events must be explained by mechanisms other than plate motions and their associated phenomena. (See PLATE TECTONICS for further information.)

Aftershocks, foreshocks, and swarms. Usually a major or even moderate earthquake of shallow focus is followed by many lesser earthquakes close to the original source

region. This is to be expected because the fault rupture producing a major earthquake does not relieve all of the accumulated strain energy at once. Furthermore, this dislocation is liable to cause an increase in the stress and strain at a number of places in the vicinity of the focal region, bringing crustal rocks at certain points close to the stress at which fracture occurs. In some cases the frequency of aftershocks may be for a time as high as 1,000 or more a day.

Sometimes a large earthquake is followed by another at approximately the same focus within an hour or perhaps a day. An extreme case of this is multiple earthquakes. In most instances, however, the first principal earthquake of a series is much more energetic than the aftershocks. In general, the number of aftershocks per day decreases with increasing time. The aftershock frequency is roughly inversely proportional to the time since the occurrence of the largest earthquake of the series.

Aftershock frequency

Most major earthquakes occur without detectable warning from less energetic precursor earthquakes, but some principal earthquakes are preceded by foreshocks. In another pattern of occurrence, large numbers of small earthquakes occur in a region over an interval of time that may extend to some months without a major earthquake occurring. In the Matsushiro region of Japan, for instance, there occurred between August 1965 and 1967 a series of hundreds of thousands of earthquakes, some sufficiently strong (up to local magnitude 5) to cause property damage but no casualties. The maximum frequency was 6,780 small earthquakes on April 17, 1966. Such series of earthquakes are called earthquake swarms. Earthquakes associated with volcanic activity often occur in swarms, but swarms also have been observed in many nonvolcanic regions.

Extraterrestrial seismic phenomena. Space vehicles have carried equipment to the surface of the Moon and Mars with which to record seismic waves, and seismologists on Earth have received telemetered signals from seismic events in both cases.

By 1969 seismographs had been placed at six sites on the Moon during the U.S. Apollo missions. Recording of seismic data ceased in September 1977. The instruments detected between 600 and 3,000 moonquakes during each year of their operation, though most of these seismic events were very small. The ground noise on the lunar surface is low compared with that of the Earth so that the seismographs could be operated at very high magnifications. Because there was more than one station on the Moon, it was possible to use the arrival times of P and S waves at the lunar stations from the moonquakes to determine foci in the same way as is done on the Earth.

Moonquakes are of three types. First, there are the events caused by the impact of lunar modules, booster rockets, and meteorites. The lunar seismograph stations were able to detect meteorites hitting the Moon's surface more than 1,000 kilometres away. The two other types of moonquakes had natural sources in the Moon's interior: they presumably resulted from rock fracturing, as on Earth. The most common type of natural moonquake had deep foci, at depths of 600 to 1,000 kilometres; the less common variety had shallow focal depths.

Causes of moonquakes

Seismological research on Mars has been less successful. Only one of the seismometers carried to the Martian surface by the U.S. Viking landers during the mid-1970s remained operational. Perhaps only one marsquake was detected in 546 Martian days.

SIZE, ENERGY, AND FREQUENCY OF EARTHQUAKES

Microseisms

As noted earlier, small ground motions known as microseisms are commonly recorded by seismographs. These weak wave motions are not generated by earthquakes, and they complicate accurate recording of the latter. They, however, are of scientific interest because their form is related to the Earth's surface structure.

Some microseisms have local causes, as, for example, those due either to traffic or machinery, or to local wind effects and storms. Another class of microseisms exhibits features that are very similar to those on records traced at earthquake observatories distributed over a wide area. The features include approximately simultaneous occurrence of maximum amplitudes and similar wave frequencies at all the observatories concerned. These microseisms may persist for many hours and have more or less regular periods of about five to eight seconds. The largest amplitudes of such microseisms are on the order of 10^{-3} centimetres and occur in coastal regions. The amplitudes also depend to some extent on local geological structure. There is a fair correlation between the size of microseisms and the occurrence of stormy weather conditions in some adjacent region.

Some microseisms are generated by the action of rough surf against an extended steep coast, while others are produced when large standing water waves are formed at sea. The period of the latter type of microseism is half that of the standing wave.

Earthquake magnitude. Because the size of earthquakes varies enormously it is necessary for purposes of relative comparison to compress the range of wave amplitudes measured on seismograms by means of a mathematical device. In 1935 the American seismologist Charles F. Richter set up a "magnitude scale of earthquakes" as the logarithm to base 10 of the maximum seismic wave amplitude (in thousandths of a millimetre) recorded on a standard seismograph (the Wood–Anderson torsion pendulum seismograph) at a distance of 100 kilometres from the earthquake epicentre. Reduction of amplitudes observed at various distances to the amplitudes expected at the standard distance of 100 kilometres is made on the basis of empirical tables. Richter magnitudes M_L are computed on the assumption that the ratio of the maximum wave amplitudes at two given distances is the same for all earthquakes considered and is independent of azimuth.

Richter magnitude scale

Richter first applied his magnitude scale to shallow-focus earthquakes recorded within 600 kilometres of the epicentre in the southern California region. Later, additional empirical tables were set up, whereby observations made at distant stations and on seismographs other than the standard type could be used. Empirical tables were extended to cover earthquakes of all significant focal depths and to enable independent magnitude estimates to be made from body- and surface-wave observations.

At the present time, a number of different magnitude scales are used by scientists and engineers as a measure of the relative size of an earthquake. The P-wave magnitude (m_b), for one, is defined in terms of the amplitude of the P wave recorded on a standard seismograph. Similarly, the surface-wave magnitude (M_s) is defined in terms of the logarithm of the maximum amplitude of the ground motion for surface waves with a wave period of 20 seconds. Taken as such, a magnitude scale has no lower or upper limit. Sensitive seismographs can record earthquakes with magnitudes of negative value and have recorded magnitudes up to about 9.0. (The 1906 San Francisco earthquake, for example, had a Richter magnitude of 8.25.)

There is, in effect, no direct mechanical basis for magnitude. Rather, it is an empirical parameter analogous to stellar magnitude. In modern practice, a more soundly based mechanical measure of earthquake size is used—namely, the seismic moment (M_0). Such a parameter is related to the angular leverage of the forces that produce the slip on the causative fault. It can be calculated both from recorded seismic waves and from field measurements of the size of the fault rupture. Consequently, seismic moment provides a more uniform scale of earthquake size. Still another magnitude currently in use is called moment magnitude (M_w). It is proportional to the logarithm of the seismic moment. Given the above definitions, the great Alaska earthquake of 1964 had the values $M_s = 8.4$, $M_0 = 820 \times 10^{27}$ dyne centimetres, $M_w = 9.2$.

Seismic moment as a measure of earthquake size

Energy and frequency of occurrence. Energy in an earthquake passing a particular surface site can be calculated directly from the recordings of strong ground motion, which is given as ground velocity. Such recordings indicate an energy rate of 10^5 watts per square metre near a moderate-sized earthquake source. The total power output of a rupturing fault in a shallow earthquake is on the order of 10^{14} watts compared with the 10^5 watts generated in rocket motors.

The magnitude M_s has also been connected with the energy E_s of an earthquake by empirical formulas. These give $E_s = 6.3 \times 10^{11}$ and 1.4×10^{25} ergs for earthquakes of $M_s = 0$ and 8.9, respectively. A unit increase in M_s thus corresponds to a 32-fold increase in energy. Negative magnitudes correspond to the smallest instrumentally recorded earthquakes, a magnitude of 1.5 to the smallest felt earthquakes and one of 3 to any shock felt at a distance of up to 20 kilometres. Earthquakes of magnitude 5.0 cause light damage near the epicentre; those of 6 are destructive over a restricted area; and those of 7.5 are at the lower limit of major earthquakes.

Total annual energy output

The total annual energy released in all earthquakes is about 10^{25} ergs, corresponding to a rate of work between 10,000,000 and 100,000,000 kilowatts. This is on the order of 0.001 of the annual amount of heat escaping from the Earth's interior. Ninety percent of the total seismic energy comes from earthquakes of magnitude 7.0 and higher—*i.e.*, those whose energy is on the order of 10^{23} ergs or more.

There also are empirical relations for the frequencies of earthquakes of various magnitudes. Suppose N to be the average number of shocks per year for which the magnitude lies in the range $M_s \pm \triangle M_s$. Then $\log_{10} N = a - bM_s$ fits the data well both globally and for particular regions; *e.g.*, for shallow earthquakes worldwide: $a = 6.7$, $b = 0.9$ when $M_s > 6.0$. The frequency for larger earthquakes therefore increases by a factor of about 10 when the magnitude is diminished by one unit. The increase in frequency with reduction in M_s falls short, however, of matching the decrease in the energy E. Thus larger earthquakes are overwhelmingly responsible for most of the total seismic energy release. The number of earthquakes per year with $m_b > 4.0$ may reach 20,000.

EARTHQUAKE PREDICTION

Observation and interpretation of precursory phenomena. The search for periodic cycles in earthquake occurrence is an old one. Generally, periodicities in time and space for major earthquakes have not been widely detected or accepted. One problem is that earthquake catalogs are not homogeneous in their selection and reporting. The most extensive catalog of this kind comes from China and begins about 700 BC. The catalog contains some information about 1,000 destructive earthquakes. The sizes of these earthquakes have been assessed from the reports of damage, intensity, and shaking.

Another approach to the statistical occurrence of earthquakes involves the postulation of trigger forces that initiate the rupture. Such forces have been attributed, for example, to severe weather conditions, volcanic activity, and tidal forces. Usually correlations are made between the physical phenomena assumed to provide the trigger and the repetition of earthquakes. Inquiry must always be made to discover whether a causative link is actually present. No trigger mechanism, at least for moderate to large earthquakes, has been found that satisfies the various criteria necessary to establish a clear physical connection.

Statistical methods also have been tried with populations of regional earthquakes. It has been suggested that the slope b of the regression line between the number of earthquakes and the magnitude, mentioned in the previous section, for a region may change characteristically with time. Specifically, the b value for the population of foreshocks of a major earthquake may be significantly smaller than the mean b value for the region averaged over a long interval of time.

For prediction of the time of earthquake occurrence, a proposal is that precursory changes in a region will cause the velocity of seismic waves through the region to change. Thus, if appropriate travel-time residuals are plotted as a function of time, fluctuations will provide a forewarning.

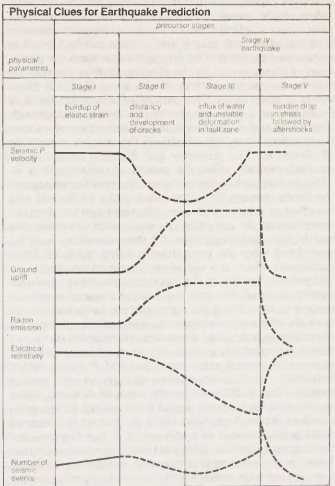
The elastic rebound theory for the occurrence of earthquakes described earlier allows rough prediction of large shallow earthquakes. H.F. Reid gave, for example, a crude forecast of the next great earthquake near San Francisco. (The theory also predicted of course that the place would be along the San Andreas or an associated fault.) The geodetic data indicated that during an interval of 50 years relative displacements of 3.2 metres had occurred at distant points across the fault. The maximum elastic-rebound offset along the fault in the 1906 earthquake was 6.5 metres. Therefore, $(6.5/3.2) \times 50$, or about 100 years, would again elapse before sufficient strain accumulated for the occurrence of an earthquake comparable to that of 1906. The premises are that the regional strain will grow uniformly and that various constraints have not been altered by the great 1906 rupture itself (*e.g.*, by the onset of slow fault slip).

For many years prediction research has been influenced by the basic argument that strain accumulates in the rock masses in the vicinity of a fault and results in crustal deformation. Deformations have been measured in the horizontal direction along active faults (by trilateration and triangulation) and in the vertical direction by precise leveling and tiltmeters. Some investigators believe that changes in groundwater level occur prior to earthquakes; variations of this sort have been reported mainly from China. It should be noted that water levels in wells respond to a complex array of factors such as rainfall; thus, if changes in water level are to be studied in relation to earthquakes, such factors need to be removed.

The theory of dilatancy of rock prior to rupture occupies a central position in recent discussions of premonitory phenomena of earthquakes. It is based on the observation that many solids exhibit dilatancy (*i.e.*, an increase in volume) during deformation. For earthquake prediction, the significance of dilatancy is its effects on various measurable quantities of the Earth's crust, such as seismic velocities, electric resistivity, and ground and water levels. The consequences of dilatancy for earthquake prediction are summarized in the Table. The best studied consequence is the effect on the seismic velocities. The influence of internal cracks and pores on the elastic properties of rocks can be clearly demonstrated in laboratory measurements of those properties as a function of hydrostatic pressure. In the case of saturated rocks, experiments predict—for shallow earthquakes—that dilatancy occurs as a portion of the crust is stressed to failure, causing a decrease in the velocities of seismic waves. Recovery of velocity is brought about by subsequent rise of pore pressure of water. The

Importance of dilatancy for earthquake prediction

Physical Clues for Earthquake Prediction



Source: Predicting Earthquakes, National Academy of Sciences, 1976

rise of pore pressure also has the effect of weakening the rock and enhancing fault slip.

Strain buildup in the focal region may have significant effects on other observable properties, including electrical conductivity and gas concentration. Because the electrical conductivity of rocks depends largely on interconnected water channels within the rocks, resistivity may increase before the cracks become saturated. As pore fluid is expelled from the closing cracks, the local water table would rise and concentrations of gases such as radioactive radon would increase.

Geological methods of extending the seismicity record back from the present also are being explored. Field studies indicate that the sequence of surface ruptures along major active faults associated with large earthquakes can sometimes be constructed. Liquefaction effects preserved in beds of sand and peat may provide evidence—when radiometric dating methods are used—for large "paleoearthquakes" extending back for more than 1,000 years.

Less well-grounded precursory phenomena, particularly earthquake lights and animal behaviour, sometimes draw more public attention than those discussed above. Many reports of unusual lights in the sky and abnormal animal behaviour preceding earthquakes are known to seismologists, mostly in anecdotal form. Both these phenomena are usually explained in terms of a release of gases prior to earthquakes and electric and acoustic stimuli of various types. At present there is no definitive experimental evidence to support claims that animals sometimes sense the coming of an earthquake.

Methods of reducing earthquake hazards. Considerable work has been done in seismology to explain the characteristics of the recorded ground motions in earthquakes. Such knowledge is needed to predict ground motions in future earthquakes so that earthquake-resistant structures can be designed. Although earthquakes cause death and destruction through such secondary effects as landslides, tsunamis, fires, and fault rupture, the greatest losses—both in lives and property—result from the collapse of man-made surface and subsurface structures during the violent shaking of the ground. Accordingly, the most effective way to mitigate the destructiveness of earthquakes from an engineering standpoint is to design and construct structures capable of withstanding strong ground motions.

Most elastic waves recorded close to an extended fault source are complicated and difficult to interpret uniquely. Understanding such near-source motion can be viewed as a three-part problem. The first part stems from the generation of elastic waves by the slipping fault as the moving rupture sweeps out an area of slip along the fault plane within a given time. The pattern of waves produced is dependent on a finite number of parameters, such as fault dimension and rupture velocity. Elastic waves of various types radiate from the vicinity of the moving rupture in all directions. The geometry and frictional properties of the fault critically affect the pattern of radiation from it.

The second part of the problem concerns the passage of the waves through the intervening rocks to the site and the effect of geological studies. The third part involves the conditions at the recording site itself, such as topography and highly attenuating soils. All these questions must be considered when an evaluation is being made of likely earthquake effects at a site of any proposed structure.

Experience has shown that accelerograms have a variable pattern in detail, but most have regular shapes in general (except in the case of strongly multiple earthquakes). An example of strong horizontal shaking of the ground (acceleration, velocity, and displacement) recorded during an actual earthquake is given in Figure 4. There is an initial segment of motion made up mainly of *P* waves, which frequently manifest themselves strongly in the vertical motion. This is followed by the onset of *S* waves, often associated with a longer period pulse related to the near-site fault slip or fling. After the *S* onset there is enhanced shaking that consists of a mixture of *S* and *P* waves, but the *S* motions become dominant as the duration increases. Later, in the horizontal component, surface waves dominate, mixed with some *S*-body waves. Depending on the distance of the site from the fault and the structure of the

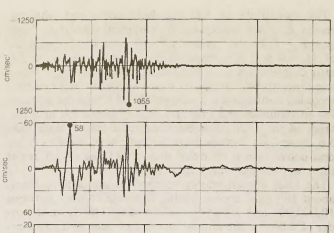


Figure 4: Recording of the San Fernando earthquake, near Pacoima Dam, California, 1971, showing (top) ground acceleration, (centre) velocity, and (bottom) displacement.

intervening rocks and soils, surface waves are spread out into long trains.

Earthquake risk maps

In many areas seismic expectancy maps or risk maps are now available for planning purposes. The anticipated intensity of ground shaking is represented by a number called the effective peak acceleration (EPA).

In order to avoid weaknesses found in earlier earthquake risk maps, the following general principles are usually adopted today: (1) the map should take into account not only the size but also the frequency of earthquakes; (2) the broad regionalization pattern should use as a data base historical seismicity, major tectonic trends, acceleration attenuation curves, and intensity reports; (3) regionalization should be defined by means of contour lines with design parameters referred to ordered numbers on neighbouring contour lines (this procedure minimizes sensitivity concerning the exact location of boundary lines between separate zones); (4) the map should be simple and not attempt to microzone the region; and (5) the mapped contoured surface should not contain discontinuities, so that the level of hazard progresses gradually and in order across any profile drawn on the map.

Developing structural designs that are able to resist the forces generated by seismic waves can be achieved either by following building codes based on risk maps or by appropriate methods of analysis. Many countries reserve theoretical structural analyses for the larger, more costly or critical buildings to be constructed in seismically active regions, while simply requiring that ordinary structures conform to local building codes. Economic realities usually determine the goal, not of preventing all damage in all earthquakes, but of minimizing damage in moderate, more common earthquakes and ensuring no major collapse at the strongest intensities. An essential part of what goes into engineering decisions on design and into the development and revision of earthquake-resistant design codes is therefore seismological, involving measurement of strong seismic waves, field studies of intensity and damage, and the probability of earthquake occurrence.

EXPLORATION OF THE EARTH'S INTERIOR WITH SEISMIC WAVES

Seismological methods and earthquake tomography. Seismological data on the Earth's deep structure come from several sources. These include *P* and *S* waves in earthquakes and nuclear explosions, the dispersion of surface waves from distant earthquakes, and vibrations of the whole Earth from large earthquakes.

One of the major aims of seismology is to infer the minimum set of properties of the Earth's interior that will explain recorded wave trains in detail. Notwithstanding the tremendous progress made in the exploration of the

Earth's deep structure during the first half of the 20th century, realization of this goal was severely limited until the 1960s because of the laborious effort required to evaluate theoretical models and to process the large amounts of seismological data recorded. The application of high-speed computers with their enormous storage and rapid retrieval capabilities opened the way for major advances in both theoretical work and data handling.

Since the mid-1970s researchers have studied realistic models of the Earth's structure that include continental and oceanic boundaries, mountains, and alluvial valleys rather than simple structures such as those involving variation only with depth. They also have resorted to statistical analyses that entail the simultaneous analyses of worldwide recordings of earthquake waves. In addition, various developments have benefited observational seismology. For example, the implications of seismic exploratory techniques developed by the petroleum industry (*e.g.*, seismic reflection) have been recognized and the procedures adopted. (For a discussion of these techniques, see EXPLORATION: *Exploration of the Earth's surface and interior.*) Equally significant has been the application of graphical methods to the exploration of the Earth's deep structure. This has been made possible by the development of minicomputers and microprocessors with peripheral display equipment.

Principal method for determining the Earth's deep structure

The major method for determining the structure of the Earth's deep interior is the detailed analysis of seismograms of seismic waves. (It is of interest that such earthquake readings also provide close estimates of wave velocities, density, and elastic and inelastic parameters in the Earth.) The primary procedure is to measure the travel times of various wave types, such as *P* and *S*, from their source to the recording seismograph. First, however, identification of each wave type with its ray path through the Earth must be made.

In Figure 5 seismic rays for many paths of *P* and *S* waves leaving the earthquake focus *F* are shown. Rays corresponding to waves that have suffered reflection at the Earth's outer surface (or possibly at one of the interior discontinuity surfaces) are denoted as *PP, PPP, SS, SSS, PS, SP, PPS*, etc. For example, *PS* corresponds to a wave that is of *P* type before surface reflection and of *S* type afterward. In addition, there are rays such as *pPP, sPP*, and *sPS*, the symbols *p* and *s* corresponding to an initial ascent to the outer surface as *P* or *S* waves, respectively. *PdP* is the *P* wave reflected from a discontinuity depth *d* kilometres in the upper part of the Earth.

An especially important class of rays is associated with a discontinuity surface that occurs at a depth of about 2,900 kilometres below the outer surface separating the central core of the Earth from the mantle. The symbol *c* is used to indicate an upward reflection at this discontinuity. Thus if a *P* wave travels down from a focus to the discontinuity surface in question, the upward reflection into an *S* wave is recorded at an observing station as the ray *PcS* and similarly with *PcP, ScS*, and *ScP*. The symbol *K* is used to denote the part (of *P* type) of the path of a wave that passes through the central core. Thus, the ray *SKS* corresponds to a wave that starts as an *S* wave, is refracted into the central core as a *P* wave, and is refracted back into the mantle wherein it finally emerges as an *S* wave. Such rays as *PKKP* correspond to waves that have suffered an internal reflection at the boundary of the central core.

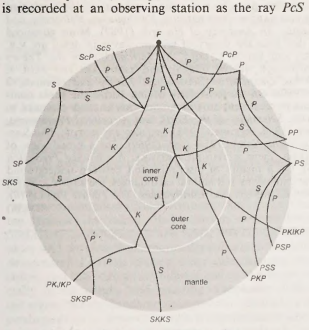


Figure 5: Seismic ray types in Earth's interior from earthquake at *F*.

The discovery of the existence of an inner core in 1936 by the Danish seismologist Inge Lehmann made it necessary to introduce additional basic symbols. For paths of waves inside the central core, the symbols *i* and *I* are used analogously to *c* and *K* for the whole Earth; therefore *i* indicates reflection upward at the boundary between the outer and inner portions of the central core, and *I* corresponds to the part (of *P* type) of the path of a wave that lies inside the inner portion. Thus, for instance, discrimination needs to be made between the rays *PKP, PKiKP*, and *PKIKP*. The first of these corresponds to a wave that has entered the outer portion of the central core but has not reached the inner portion; the second to one that has been reflected upward at the boundary between the two portions; and the third to one that has penetrated into the inner portion.

By combining the symbols *p, s, P, S, c, K, i, I*, and *d* in various ways, notation is developed for all the main rays associated with body earthquake waves. The symbol *J* has been introduced to correspond to *S* waves in the inner core, should evidence be found for such waves.

Finally, the use of times of travel along rays to infer hidden structure is analogous to the use of X rays in medical tomography. The method involves reconstructing an image of internal anomalies from measurements made at the outer surface. Nowadays, hundreds of thousands of travel times of *P* and *S* waves are available in earthquake catalogs for the tomographic imaging of the Earth's interior and the mapping of internal structure.

Structure of the Earth's interior. Studies with earthquake recordings have given a picture inside the Earth of, on average, a solid but layered mantle about 2,900 kilometres thick, which in places lies within 10 kilometres of the surface under the oceans. The thin rock layer surrounding the mantle is the crust, whose lower boundary is called the Mohorovičić Discontinuity.

In normal continental regions of 30- to 40-kilometre thickness, there is usually a superficial low-velocity sedimentary layer underlain by a zone in which seismic velocity increases with depth. This may be followed by a layer in which *P*-wave velocities in some places fall from 6 to 5.6 kilometres per second. The middle part of the crust is characterized by a heterogeneous zone with *P* velocities of nearly 6 to 6.3 kilometres per second. The lowest layer of the crust (about 10 kilometres thick) has significantly higher *P* velocities ranging up to nearly 7 kilometres per second.

In the deep ocean, under a sedimentary layer of about one-kilometre thickness, the lower layer of the thin oceanic crust is inferred to consist of basalt, which formed where extrusions of basaltic magma at mid-ocean ridges have been added to the upper part of lithospheric plates as they spread away from the ridge crests. This crustal layer cools as it moves away from the ridge crest, and its seismic velocities increase correspondingly.

Below the mantle lies a 2,255-kilometre-thick shell, which seismic waves show to have the properties of a liquid. At the very centre of the planet is a separate solid core, with a radius of 1,216 kilometres. Recent work with observed seismic waves has revealed fine structural details within the main shells inside the Earth, especially the crust and lithosphere. These regional variations are important in explaining the dynamic history of the planet.

Long-period oscillations of the globe. Sometimes earthquakes are large enough to cause the whole Earth to ring like a bell. The deepest tone of vibration of the planet is one of 54 minutes. Knowledge of these vibrations has come from a remarkable extension in the range of periods of ground movements that can be recorded by very long-period seismographs, thus allowing the interval in earthquake wave periods to be filled in: from ordinary *P* waves

with periods of a few seconds to vibrations with periods on the order of 12 and 24 hours such as those that occur in Earth tidal movements.

The measurements of the vibrations of the whole Earth provide important additional data on the properties of the interior of the planet. It should be emphasized that these free vibrations are set up by the energy release of the earthquake source but continue for many hours and sometimes even days. For an elastic sphere like the Earth two types of vibrations are known to be possible. In one type, called *S* modes or spheroidal vibrations, the motions of the elements of the sphere have components along the radius as well as along the tangent. In the second type, which are designated as *T* modes or torsional vibrations, there are shear but no radial displacements. The nomenclature is ${}_nS_\ell$ and ${}_nT_\ell$, where the letters n and ℓ are related to the surfaces in the vibration at which there is zero motion. Four examples are illustrated in Figure 6. The suffix n gives a count of the number of internal zero-motion (nodal) surfaces and the suffix ℓ indicates the number of surface nodal lines.

Spheroidal and torsional vibrations

Several hundred types of *S* and *T* vibrations have been identified and the associated periods measured. In a smaller number of cases, the amplitude of the ground motion in the vibrations has been determined for particular earthquakes, and, more importantly, the attenuation of each component vibration has been measured. The measure of this decay constant is called the quality factor *Q*. The greater the value of *Q*, the less is the wave or vibration damping. Typically, for ${}_0S_{10}$ and ${}_0T_{10}$, the *Q* values are about 250.

The rate of decay of the vibrations of the whole Earth with the passage of time can be seen in Figure 7, where they appear superimposed for 20 hours of the 12-hour tidal deformations of the Earth. At the bottom of Figure 7, these vibrations have been split up into a series of peaks, each with a definite frequency, like the spectrum of light. Such a spectrum indicates the relative amplitude of each harmonic present in the free oscillations. If the physical properties of the Earth's interior were known, all these individual peaks could be calculated directly. Instead, the internal structure must be estimated from the observed peaks.

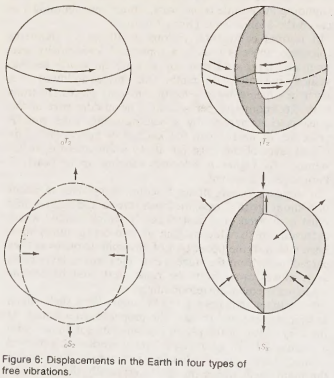


Figure 6: Displacements in the Earth in four types of free vibrations.

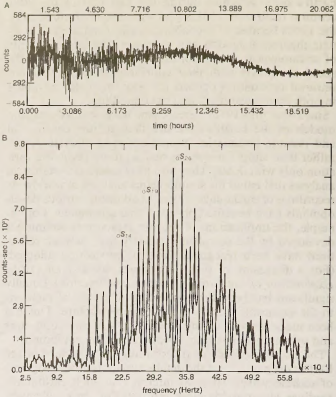


Figure 7: (A) Recorded ground motion for 20 hours at Whiskeytown, Calif., in the large Indonesian earthquake, 1977. (B) Frequency spectrum of the Earth's oscillations from the record in A above.

Recent research has shown that observations of long-period oscillations of the Earth discriminate fairly finely between different Earth models. In applying the observations to improve the resolution and precision of such representations of the planet's internal structure, a considerable number of Earth models are set up and all the periods of their free oscillations are computed and checked against the observations. Models can then be steadily eliminated until only a small range remains. In practice, the work starts with existing models; efforts are made to amend them by successive steps until full compatibility with the observations is achieved within the uncertainties of the observations. Even so, the resulting computed Earth structure is not a unique solution to the problem.

BIBLIOGRAPHY. The subject of earthquakes is dealt with mainly in books on seismology. Recommended elementary texts are BRUCE A. BOLT, *Earthquakes: A Primer* (1978), and *Inside the Earth: Evidence from Earthquakes* (1982); and CHARLES F. RICHTER, *Elementary Seismology* (1958). Interesting discussions are also given by G.A. EBY, *Earthquakes* (1967, reissued 1980); and KARL V. STEINBRUGGE, *Earthquakes, Volcanoes, and Tsunamis: An Anatomy of Hazards* (1982). More advanced texts that treat the theory of seismic waves in detail are K.E. BULLEN and BRUCE A. BOLT, *An Introduction to the Theory of Seismology*, 4th ed. (1985); and KEIITI AKI and PAUL J. RICHARDS, *Quantitative Seismology: Theory and Methods*, 2 vol. (1980). On the seismicity of the Earth, the most comprehensive treatment is still B. GUTENBERG and CHARLES F. RICHTER, *Seismicity of the Earth and Associated Phenomena*, 2nd ed. (1954, reprinted 1965); see also J.P. ROTHÉ, *The Seismicity of the Earth, 1953–1965* (1969). For a broad view of prediction, there is TSUNEJI RIKITAKE, *Earthquake Prediction* (1976). On a history of discrimination between underground nuclear explosions and natural earthquakes, see BRUCE A. BOLT, *Nuclear Explosions and Earthquakes: The Parted Veil* (1976).

(B.A.B.)

East Asian Arts

The term East Asia, as used in this article, comprises China, Korea, and Japan. Some studies of East Asia also include the cultures of the Indochinese peninsula and adjoining islands, as well as Mongolia to the north. The logic of this occasional inclusion is based on a strict geographic definition as well as a recognition of common bonds forged through the acceptance of Buddhism by many of these cultures. China, Korea, and Japan, however, have been uniquely linked for several millennia by a common written language and by broad cultural and political connections that have ranged in spirit from the uncritically adorational to the contentious.

From ancient times, China has been the dominant and referential culture in East Asia. Although variously developed Neolithic cultures existed on the Korean Peninsula and on the Japanese archipelago, archaeological evidence in the form of worked stone and blades from the Paleolithic and Neolithic periods suggests an exchange between the early East Asian cultures and the early introduction of Chinese influence. This cultural interaction was facilitated in part by land bridges that connected Japan with the continent.

Significant developments in the production of earthenware vessels from about 10,000 BC in Japan (thus far, the world's earliest dated pottery) and from approximately 3500 BC in Korea are well documented. They reveal a rich symbolic vocabulary and decorative sense as well as a highly successful union of function and dynamic form. These types of vessels chronicle the increasing needs for storage as there was a gradual societal transformation from nomadic and foraging cultures to more sedentary crop-producing cultures. There were pottery-dominant cultures in China as well. The painted (*c.* 5000 BC) and black (*c.* 2500 BC) earthenware are the best known.

As Korea and Japan continued in various Neolithic phases, developments in China from approximately 2000 BC were far more complex and dramatic. Archaeological evidence firmly corroborates the existence of an emerging bronze culture by approximately 2000 BC. This culture provided the base for Shang dynasty (approximately from the 16th to the 11th century BC) culture, which witnessed extraordinary developments in the production of bronze, stone, ceramic, and jade artifacts as well as the development of a pictograph-based written language. Bronze production and the expansion of rice cultivation gradually appeared in Korea from approximately 700 BC and then slightly later in Japan. While no single political event seemed to further the transmission of Chinese cultural elements to Korea and Japan, clearly the expansionist policies of the rulers of the Han dynasty (206 BC–AD 220) stimulated what had been a gradual assimilation of Chinese cultural elements by both Korea and Japan. Indicatively, it is from this period that Chinese documentation of legation visits to Japan provide the first written records describing the structure of Japanese society.

The cultures of China, Korea, and Japan went on, from this period of interaction during the Han dynasty, to develop in quite distinctive ways. China, for example, experienced two major dynasties, the Han and the T'ang (618–907), that were truly international in scope and easily rivaled contemporary Mediterranean powers. In succeeding dynasties, including rule by foreign invaders from the north, the development of the visual arts continued to explore and develop the basic media for which the Chinese demonstrated special affinity: clay, jade, lacquer, bronze, stone, and the various manifestations of the brush, especially in calligraphy and painting. Emphases shifted, as did styles, but the fundamental symbolic vocabulary and a predisposition to renew through reinterpretation and reverence of the past was characteristic not only of Chinese but of all the East Asian arts.

Korea's pivotal location gave it particular strategic value and thus made it the target of subjugation by a stronger China and Japan. But Korea strove to maintain its own identity and to prevent China and Japan from exercising control over more than a small portion of the peninsula. National contributions to the larger aesthetic culture of East Asia included unequaled mastery of goldsmithing and design as well as a ceramic tradition that included delicate celadon ware and a vigorous folk ware that inspired generations of Japanese tea masters. Indeed, Korea was a primary conduit of continental culture to the Japanese in many areas of visual expression, including metalwork, painting, and ceramics.

In the late 13th century, Mongol forces made two unsuccessful attempts at invading the Japanese islands, and the country was spared occupation by a foreign power until well into the 20th century. This unusual condition of comparative isolation provided Japanese cultural arbiters with a relative freedom to select or reject outside styles and trends. Nevertheless, Chinese art's highly developed, systematic forms of expression, coupled with its theoretical basis in religion and philosophy, proved enormously forceful, and Chinese styles dominated at key junctures in Japanese history. The reception and assimilation of outside influence followed by a vigorous assertion of national styles thus characterized the cycle of Japanese cultural development. In addition to distinctive reinterpretations of Chinese ink monochrome painting and calligraphy, an indigenous taste for the observation and depiction of human activity and an exquisitely nuanced sense of design are readily apparent in most areas of Japanese visual expression, none more so than in narrative painting and in the woodblock print.

The elements and tendencies common to the Chinese, Korean, and Japanese cultures are vast, but two kinds of visual expression are especially important: a strong affinity for the clay-formed vessel and calligraphic expression through the ink-charged brush. Vigorous, subtle, and technically sophisticated expressions ranging from Neolithic earthenware to celadon and glazed enamelware were both integral to daily life and prized by connoisseurs who judged ceramics by an elaborate code of appreciation. Increasingly abstracted forms of pictographs provided a means of writing that was image-based; characters formed by the brush could be normative but also offered infinite possibilities for personal expression through ink modulation and idiosyncratic gesture. Although Korea and Japan later developed phonetic syllabaries, the visual language of the educated continued to be based on the ancestral Chinese form. The meanings of words, phrases, or whole texts could be expanded or nuanced by their visual renderings. Painting was derivative from calligraphy, and implicit in painting skill was a preceding mastery of the brush-rendered calligraphic line. As a consequence, calligraphy was unequaled as the major element in the transmission of cultural values, whether as information or as aesthetic expression.

The influence of Buddhism, a force which was initially foreign to East Asia, also should not be underestimated. Emerging from India and Central Asia in the first century after nearly 500 years of development on the subcontinent, Buddhism offered a convincing universalist system of belief that assimilated and frequently gave visual expression to indigenous religions. By the 5th century AD, a Chinese dynastic line had adopted Buddhism as a religion of state. While individual rulers, courts, or dynasties at times propelled the florescence of East Asian arts, none of them equaled the patronage of Buddhism in duration, scale, and intellectual sustenance. Confucianism, Taoism, and, to a somewhat lesser degree, Shintō required expression through the arts; however, Buddhism's multiple

sects, complex iconography, and program of proselytizing made it the natural and dominant vehicle of transcultural influence in East Asia.

The unity and diversity of the three East Asian cultures are explored in greater depth in the article, which treats both the visual and the performing arts. The literatures of the respective languages are treated individually in the *Macropædia*. (J.T.U.)

For coverage of related topics in the *Macropædia* and *Micropædia*, see the *Propædia*, sections 613, 622, 624, 625, 626, 627, and 629, and the *Index*.

The article is divided into the following sections:

Visual arts 668
  Chinese visual arts 668
    General characteristics
    Stylistic and historical development
  Korean visual arts 701
    General characteristics
    Stylistic and historical development
  Japanese visual arts 709
    General characteristics
    Stylistic and historical development
Music 737
  The nature of East Asian music 737
    East Asian music vis-à-vis that of other major cultures
    Musical traits common to East Asian cultures
  The music of China 738
    Formative period
    T'ang dynasty
    Sung and Yüan dynasties
    Ming and Ch'ing dynasties
    Developments since 1911
  The music of Korea 745
    Shaman music
    Court instrumental music
    Vocal music
    Modern music
  The music of Japan 747
    Music before and through the Nara period
    The Heian period
    Kamakura, Muromachi, and Tokugawa periods
    The Meiji period and subsequent music
Dance and theatre 758
  Characteristics of East Asian dance and theatre 759
    Common traditions
    Social conditions
  The development of dance and theatre in the East Asian nations 760
    China
    Korea
    Japan
Bibliography 769

VISUAL ARTS

Chinese visual arts

The present political boundaries of China, which include Tibet, Inner Mongolia, Sinkiang, and the northeastern provinces formerly called Manchuria, embrace a far larger area of East Asia than will be discussed here. "China Proper," as it has been called, consists of 18 historical provinces bounded by the Tibetan Highlands on the west, the Gobi to the north, and Myanmar (Burma), Laos, and Vietnam to the southwest; and it is primarily the arts of this area that will be treated here.

The first communities that can be identified culturally as Chinese were settled chiefly in the basin of the Huang Ho (Yellow River). Gradually they spread out, influencing other tribal cultures, until, by the Han dynasty (206 BC–AD 220), most of China proper was dominated by the culture that had been formed in the "cradle" of northern Chinese civilization. Over this area there slowly spread a common written language, as well as a common belief in the power of heaven and the ancestral spirits to influence the living and in the importance of ceremony and sacrifice to achieve harmony among heaven, nature, and humankind. These beliefs were to have a great influence on the character of Chinese art.

Chronology of Chinese art

The Chinese themselves were among the most historically conscious of all the major civilizations and were intensely aware of the strength and continuity of their cultural tradition. They viewed history as a cycle of decline and renewal associated with the succession of ruling dynasties. Both the political fragmentation and social and economic chaos of decline and the vigour of dynastic rejuvenation could stimulate and colour important artistic developments. Thus, it is quite legitimate to think of the history of Chinese art, as the Chinese themselves do, primarily in terms of the styles of successive dynasties.

GENERAL CHARACTERISTICS

Aesthetic characteristics and artistic traditions. *Art as a reflection of Chinese class structure.* One of the outstanding characteristics of Chinese art is the extent to which it reflects the class structure that has existed at different times in Chinese history. Up to the Warring States period (475–221 BC), the arts were produced by anonymous craftsmen for the royal and feudal courts. During the Warring States and the Han dynasty, the growth of a landowning and merchant class brought new patrons; and after the Han there began to emerge the concept of "fine art" as the product of the leisure of the educated gentry, many of whom were amateur practitioners of the arts of poetry, music, calligraphy, and, eventually, painting. At this time a distinction began to arise between the lower-class professional and the elite amateur artist that was to have a great influence on the character of Chinese art in later times. Gradually one tradition became increasingly identified with the artists and craftsmen who worked for the court or sold their work for profit. Identified with another tradition, the scholarly amateurs looked upon such people with some contempt, and the art of the literati became increasingly refined and rarefied to the point that, from the Sung dynasty (960–1279) onward, an assumed awkwardness in technique was admired as a mark of the amateur and gentleman. One effect of the revolutions of the 20th century has been the breaking down of the class barriers between amateur and professional and even, during the Great Proletarian Cultural Revolution of 1966–76, an emphasis on anonymous, proletarian-made art like that of the T'ang dynasty (618–907) and earlier.

The role of calligraphy in Chinese art. Since the 3rd century AD, calligraphy, or writing as a fine art, has been considered supreme among the visual arts in China. Not only does it require immense skill and fine judgment, but it is regarded as uniquely revealing of the character and breadth of cultivation of the writer. Since the time when inscribed oracle bones and tortoise shells (China's oldest extant writing) were used for divination in the Shang dynasty (16th to 11th centuries BC), calligraphy has been associated with spiritual communication and has been viewed in terms of the writer's own spiritual attunement. To fully appreciate calligraphy, as to produce it, requires lofty personal qualities and unusual aesthetic sensitivity.

The Chinese painter uses essentially the same materials as the calligrapher—brush, ink, and silk or paper—and the Chinese judge his work by the same criteria, basically the vitality and expressiveness of the brushstroke itself and the harmonious rhythm of the whole composition. Painting in China, therefore, is essentially a linear art. The painters of most periods were concerned less with striving for originality or conveying a sense of reality and three-dimensional mass through such aids as shading and perspective, concentrating instead on transmitting to silk or paper, through the rhythmic movement of the brushstroke, an awareness of the inner life of things.

The aesthetics of line in calligraphy and painting have had a significant influence on the other arts in China.

Influence of painting and calligraphy on the other arts

In the motifs that adorn the ritual bronzes, in the flow of the drapery over the surface of Buddhist sculpture, in the decoration of lacquerware, pottery, and cloisonné enamel (wares decorated with enamel of different colours separated by strips of metal), it is the rhythmic movement of the line, following the natural movement of the artist's or craftsman's hand, that to a large extent determines the form and gives to Chinese art as a whole its remarkable harmony and unity of style. (For information about Chinese calligraphy, see the article WRITING: *East Asian calligraphy*.)

Characteristic themes and symbols. In early times this sense of attunement involved submission to the Will of Heaven through ritual and sacrifice, and it was the function of Chinese art to serve these ends. Archaic bronze vessels were made for sacrifices to heaven and to the spirits of clan ancestors, who were believed to influence the living for good if the rites were properly and regularly performed. Chinese society, basically agricultural, has always laid great stress on the need for humans to understand the pattern of nature and to live in accordance with it. The world of nature was seen as the visible manifestation of the workings of the Great Ultimate through the generative interaction of the yin-yang (female-male) dualism. As it developed, the purpose of Chinese art turned from propitiation and sacrifice to the expression of human understanding of these forces through the painting of landscape, bamboo, birds, and flowers. This might be called the metaphysical, Taoist aspect of Chinese painting.

Particularly in early times, art also had social and moral functions. The earliest paintings referred to in ancient texts depicted on the walls of palaces and ancestral halls benevolent emperors, sages, virtuous ministers, loyal generals, and their evil opposites as examples and warnings to the living. Portrait painting also had this moral function, depicting not the features of the subject so much as his character and his role in society. Court painters were called upon to depict auspicious and memorable events. This was the ethical, Confucian function of painting.

High religious art as such is foreign to China. Popular folk religion was seldom an inspiration to great works of art, and Buddhism, which indeed produced many masterpieces of a special kind, was a foreign importation.

Themes not represented in Chinese art

Among the typical themes of Chinese art there is no place for war, violence, the nude, death, or martyrdom. Nor is inanimate matter ever painted for art's sake: the very rocks and streams are felt to be alive, visible manifestations of the invisible forces of the universe. No theme would be accepted in Chinese art that was not inspiring, noble, refreshing to the spirit, or at least charming. Nor is there any place in the Chinese artistic tradition for an art of pure form divorced from content, and the Chinese cannot conceive of a work of art of which the form is beautiful while the subject matter is unedifying. In the broadest sense, therefore, all Chinese art is symbolic, for everything that is painted reflects some aspect of a totality of which the painter is intuitively aware. At the same time Chinese art is full of symbols of a more specific kind, some with various possible meanings.

General media characteristics. *Architecture.* Because the Chinese build chiefly in timber, which is vulnerable to fire and the ravages of time, very little ancient architecture has survived. The oldest datable timber building is the small main hall of the Nan-ch'an Temple, on Mount Wu-t'ai in Shansi province, built sometime before 782 and restored in that year. Brick and stone are used for defensive walls, the arch for gates and bridges, and the vault for tombs. Only rarely has the corbeled dome (in which each successive course projects inward from the course below it) been used for temples and tombs. Single-story architecture predominates throughout northern and much of eastern China, although multistory building techniques date to the late Chou dynasty (11th century–255 BC).

The basic elements in a Chinese timber building are the platform of pounded earth faced with stone or tile on which the building stands; the post-and-lintel frame (vertical posts topped by horizontal tie beams); the roof-supporting brackets and truss; and the tiled roof itself. The walls between the posts, or columns, are not load-bearing, and the intercolumnar bays (odd-numbered along the front of the building) may be filled by doors (usually doubled) or by brick or such material as bamboo wattle faced with plaster, or they may be left open to create peristyles. The flexible triangular truss is placed transverse to the front side of the building and defines a gable-type roof by means of a stepped-up series of elevated tie beams (*t'ai-liang*, for which this entire system of architecture is named); the beams are sequentially shortened and alternate with vertical struts that bear the roof purlins and the main roof beam. The flexible proportions of the gable-end framework of struts and beams permits the roof to take any profile desired, typically a low and rather straight silhouette in northern China before the Sung, and increasingly elevated and concave in the Sung, Yüan, Ming, and Ch'ing. The gable-end framework is typically moved inward in a prominent building and partially masked in a hip-and-gable (or half-hip) roof and completely masked in a full-hipped roof. The timber building is limited in depth by the span of the truss; in theory, however, it may be of any length, although it rarely exceeds 11 bays in practice.

The origin of the distinctive curve of the roof, which first appeared in China about the 6th century AD, is not fully understood, although a number of theories have been put forward.

The "pavilion concept"

In the "pavilion concept," whereby each building is conceived of as a freestanding rectilinear unit, flexibility in the overall design is achieved by increasing the number of such units, which are arranged together with open, connecting galleries around courtyards; diversity is achieved through design variations that individualize these courtyard complexes. In the private house or mansion, the grouping of halls and courtyards is informal, apart from the axial arrangement of the entrance court with its main hall facing the gateway; but in a palace such as the gigantic Forbidden City in Peking, the principal halls are ranged with their courtyards behind one another on a south-to-north axis, building up to a ceremonial climax and dying away to lesser courts and buildings to the north. Ancestral halls and temples follow the palatial arrangement. The scale of a building, the number of bays, the unit of measure used for the timbers, whether bracketing is included or not, and the type of roof (gabled, half- or full-hipped, with or without prominent decorative ridge-tiling and prominent overhang) all accord with the placement and significance of the building within a courtyard arrangement, with the relative importance of that courtyard within a larger compound, and with the absolute status of the whole building complex. The entire system, therefore, is modular and highly standardized.

The domination of the roof allows little variation in the form of the individual building; thus, aesthetic subtlety is concentrated in pleasing proportions and in details such as the roof brackets or the plinths supporting the columns. T'ang architecture achieved a "classic" standard, with massive proportions yet simple designs in which function and form were fully harmonized. Architects in the Sung dynasty were much more adventurous in playing with interlocking roofs and different levels than were their successors in later centuries. The beauty of the architecture of the Ming dynasty (1368–1644) and Ch'ing dynasty (1644–1911/12) lies rather in the lightweight effect and the richness of painted decoration.

Painting and calligraphy. The character of Chinese painting and calligraphy is closely bound up with the nature of the medium. The basic material is ink, formed into a short stick of hardened pine soot and glue, which is rubbed to the required consistency on an inkstone with a little water. The calligrapher or painter uses a pointed-tipped brush made of the hair of goat, deer, or wolf set in a shaft of bamboo. He writes or paints on a length of silk or a sheet of paper, the surface of which is absorbent, allowing no erasure or correction. He must therefore know beforehand what he intends to do, and the execution demands confidence, speed, and a mastery of technique acquired only by long practice. For example, to broaden the brush stroke, the calligrapher or painter applies downward pressure on the brush. Such subtle action of the highly flexible but carefully controlled brush tip determines the

dynamic character of the brushwork and is the primary focus of attention of both the artist and critical viewers.

In painting, colour is added, if at all, to make the effect more true to life or to add decorative accent and rarely as a structural element in the design, as in Western art. Brighter, more opaque pigments derived from mineral sources (blue from azurite, green from malachite, red from cinnabar or lead, yellow from orpiment or ochre, all produced in various intensities) are preferred for painting on silk, while translucent vegetable pigments predominate in painting on paper (indigo blue, red from safflower or madder, vegetable green, rattan and sophora plant yellow) and produce a lighter, more delicate effect.

While painting on dry plaster walls or screens is an ancient art in China, more common formats in the past millennium have been the vertical hanging scroll, perhaps derived from the Buddhist devotional banner, and the horizontal hand scroll, which may be of any length up to about 15 metres (50 feet). Other forms are fan painting and the album leaf. The artist's carefully placed signature, inscription, and seals are an integral part of the composition. In Chinese eyes a picture may gain considerably in interest and value from the colophons added by later connoisseurs on the painting itself or, in the case of a hand scroll, mounted after it. The mounting of paintings and calligraphy is a highly skilled craft and, if carefully done, will enhance the appearance of a scroll and ensure its preservation for many centuries.

Sculpture. With rare exceptions, sculptors were regarded as mere craftsmen, and very few of their names are known. Works of sculpture were created not (like paintings) as art objects in themselves but for a specific ceremonial, religious, or funerary purpose. While small figures were carved in stone in the Shang dynasty, large-scale stone sculpture is a later development, possibly stimulated by contacts with western Asia in the Han dynasty. Major works of religious sculpture did not appear until Buddhism had taken firm root in northern China after the Han dynasty.

Clay modeling and bronze casting

More truly Chinese than stone carving is the tradition of clay modeling and its derivative, bronze casting. Tomb figurines of remarkable plasticity and liveliness were made from the Ch'in dynasty (221–206 BC) through the T'ang dynasty, while some temple sculpture was carried out on a large scale in clay, using straw or cotton as a binder. The aesthetic of the fluid medium of clay modeling, which follows the natural movement of the craftsman's hand and arm, is influenced by that of painting and calligraphy, which even affected the more intractable media of stone and wood carving.

In Sung and later times, the distinctive, abstract medium of assembling stones as "false mountains" (*chia-shan*) in royal parks and scholars' gardens came to the forefront of the Chinese sculptural arts.

Jade, lacquer, textiles, and other media. Other major art forms of China include pottery, jade carving, metalwork (including gold and silver inlay and cloisonné enamel), textiles, and lacquerware. In several of these, China can claim a long priority over the rest of the world. True pottery glazes were developed in China before the end of the 2nd millennium BC and porcelain by the 6th century, more than 1,000 years before its discovery in Europe; jade carving, sericulture (the raising of silkworms), and weaving of silk go back to Neolithic times and lacquer painting to the Shang dynasty. Bronze casting, while not so ancient as that of the Middle East, reached by 1000 BC a perfection of beauty and craftsmanship not matched in the ancient Western world. In point of style, all these arts share with sculpture a debt to pictorial art and an aesthetic based on the rhythmic movement of the line.

Jade

Jade occupies a special place in Chinese artistic culture, valued as gold is in the West but hallowed with even loftier moral connotations. Because of this and the belief in its indestructibility, jade from early times was lavishly used not only for dress ornaments but also for ritual objects, both Confucian and Taoist, and for the protection of the dead in the tomb.

The jade stone used since ancient times in China is nephrite, a crystalline calcium magnesium silicate, which in its pure state is white but may be green, cream, yellow, brown, gray, black, or mottled owing to the presence of impurities, chiefly iron compounds. Generically, the Chinese used the term *yü* to cover a variety of related "jade" stones, including nephrite, bowenite, and jadeite. In the Neolithic Period, by the mid-4th millennium BC, jade from Lake T'ai (in Kiangsu province) began to be used by southeastern culture groups, while deposits along the Liao River in the northeast (called "Hsiu-yen jade," probably bowenite) were utilized by the Hung-shan culture. In historic times, China's chief source of nephrite has been the riverbeds of Yarkand and Ho-t'ien in present-day Sinkiang autonomous region in northwestern China, where jade is found in the form of boulders. Since the 18th century, China has received from northern Myanmar (upper Burma) a brilliant green jadeite (also called *feitsui*, or "kingfisher-feathers") that is a granular sodium-aluminum silicate harder than but not quite so tough as nephrite. Having a hardness like that of steel or feldspar, jade cannot be carved or cut with metal tools but has to be laboriously drilled, ground, or sawed with an abrasive paste and rotational or repetitive-motion machinery, usually after being reduced to the form of blocks or thin slabs.

The Chinese had discovered as early as the Shang dynasty that the juice of the lac tree, a naturally occurring polymer, could be used for forming hard but lightweight vessels when built up in very thin layers through the repeated dipping of a core of carved wood, bamboo, or cloth. With the addition of pigments, most commonly red and black, less frequently green and yellow, it could also be used for painting and decorating the outer layers of these vessels. Being sticky, painted lacquer must be applied slowly with the brush, giving rise to prolonged motions and fluid, often elegantly curvilinear designs. Since lacquer is almost totally impervious to water, vessels and wine cups have been excavated in perfect condition from waterlogged graves of the late-5th-century-BC Tseng state in Sui-hsien, of the 4th–3rd-century-BC Ch'u state in Chiang-ling (now Sha-shih), and of the early-2nd-century-BC Han dynasty in Ch'ang-sha. Such works ranged from large-scale coffins to bird- or animal-shaped drum stands to such daily utensils as nested toiletry boxes and food-serving implements. By the Warring States period, lacquerwork had developed into a major industry; and, being approximately 10 times more costly than their bronze equivalents, lacquer vessels came to rival bronzes as the most esteemed medium for providing offerings in ancestral ceremonies among the wealthy aristocracy.

Sericulture

It was the Chinese who first discovered that the roughly 1 kilometre (1,000 yards) of thread that constitutes the cocoon of the silkworm, *Bombyx mori*, could be reeled off, spun, and woven; and sericulture early became an important feature of Chinese rural economy. Its place in Chinese culture is indicated by the legend that it was the wife of the mythical Yellow Emperor, Huang-ti, who taught the Chinese people the art and by the fact that in historic times the empress was ceremonially associated with it. The weaving of damask probably existed in the Shang dynasty, and the 4th–3rd-century-BC tombs at Ma-shan near Chiang-ling (Hupeh province, excavated in 1982) have provided outstanding examples of brocade, gauze, and embroidery with pictorial designs as well as the first complete garments. Although transportation westward across Central Asia's trade route brought Chinese silks to many parts of the Mediterranean region, the knowledge of silk production techniques did not reach the area until the 6th century AD.

STYLISTIC AND HISTORICAL DEVELOPMENT

Formative period. The earliest evidence for art in any form in ancient China consists of crude cord-marked pottery and artifacts decorated with geometric designs found in Mesolithic sites in northern China and in the Kwangtung-Kwangsi regions. The dating for prehistoric culture in China is still very uncertain, but this material is probably at least 7,000 or 8,000 years old. The art of the Neolithic Period represents a considerable advance. The Yang-shao, or Painted Pottery, culture (named after the first Neolithic site discovered, in 1920), which had its centre around the

eastern bend of the Huang Ho, is now known to have extended across northern China and up into Kansu province. Yang-shao pottery consists chiefly of full-bodied funerary storage jars made by the coiling, or ring, method. They are decorated, generally on the upper half only, with a rich variety of geometric designs, whorls, volutes, and sawtooth patterns executed in black and red pigment with a brush whose sweeping, rhythmic lines foreshadow the free brush painting of historic periods. Some of the pottery from the village site of Pan-p'o (*c.* 4500 BC), discovered in 1953 near Sian in Shensi, bear schematized fish, bird, deer, and plant designs, which are related thematically to hunting and gathering, and what may be a human face or mask. Dating for the dominant phase of Yang-shao culture may be put roughly between 5000 and 3000 BC. Over this span of two millennia the Yang-shao culture progressed generally westward along the Huang Ho and Wei River valleys from sites in central China, Kansu province, such as Pan-p'o, to sites including Miao-ti-kou, Ma-chia-yao, Pan-shan, and Ma-ch'ang. The art produced at these villages exhibits a clear and logical stylistic evolution, leading from representational designs to linear abstraction (the latter with occasional symbolic references).

Lung-shan culture

The last major phase of the Neolithic Period is represented by Lung-shan culture, distinguished particularly by the black pottery of its later stages (*c.* 2200–1700 BC). Lung-shan is named after the site of its discovery in 1928, in Shantung province, although evidence increasingly suggests origins to the south along the China coast, in Kiangsu province. By contrast with the Yang-shao, the fully developed Lung-shan pottery is wheel-made and especially thinly potted. The finest specimens have a dark gray or black body burnished to a hard, smooth surface that is occasionally incised but never painted, giving it a metallic appearance. The occasional use of open-worked design and the simulation of lugs and folded plating all suggest the highly skilled imitation, in an inexhaustible medium, of valuable copper wares which, although no longer extant, heralded the transition from a lithic to a metallic culture. At this point, the superior calibre of Chinese ceramics was first attained.

In Yang-shao pottery, emphasis was on funerary wares. The delicate potting of the Lung-shan ware and the prevalence of offering stands and goblets suggest that these vessels were made not for burial but for sacrificial rites connected with the worship of ancestral spirits. Ritual vessels, oracle bones (used by shamans in divination), ceremonial jade objects and ornaments, and architecture (pounded-earth foundations, protective city walls, rectilinear organization) reflect an advanced material culture on the threshold of the Bronze Age. This culture continued in outlying areas long after the coming of bronze technology to the central Honan–Shensi–southern Shansi region.

The earliest examples of jade from the lower Yangtze River region appear in the latter phases of Ma-chia-pang culture (*c.* 5100–3900 BC) and continue into the 4th–3rd millennia BC in the Sung-tse and Ch'ing-lien-kang cultures of that region. Remarkably sophisticated jade pieces appear after 2500 BC in the Liang-chu culture of southern Kiangsu and northern Chekiang provinces (*c.* 3400–2200 BC), many with an apparent lack of wear and practical usage that suggests a primarily ceremonial function. These include the first examples of the flat, perforated *pi* disk, which became the symbol of heaven in later times, and of the *ts'ung*, a tube with a square exterior and cylindrical hollow centre. These two items remained part of the Chinese Imperial paraphernalia until the early 20th century. The precise meaning of the *ts'ung* and its possible association with astronomical sighting or geomantic site selection or its conjunction of yin (square, earth, female) and yang (circular, heaven, male) features remains unclear. Also present at this time, in the Liang-chu and the Shantung province Lung-shan cultures, are ceremonial *kuei* and *chang* blades and axes, as well as an increasing variety of ornamental arc-shaped and circular jade pendants, necklaces, and bracelets (often in animal form), together with the significant appearance of mask decoration, all of which link these Neolithic jades to those of the subsequent Shang period.

The *pi* and the *ts'ung*

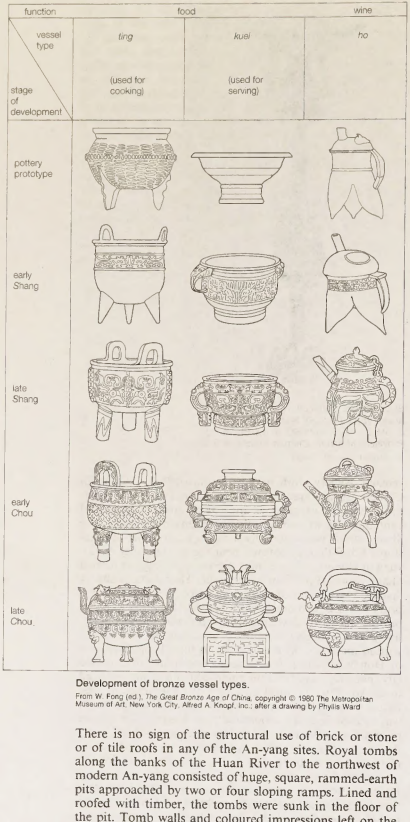


Black pottery stem cup, Neolithic Lung-shan phase, from Jih-chao, Shantung province, *c.* late 3rd millennium BC. In the Shantung Provincial Museum, Chi-nan. Height 26.5 cm.
Wang Lu/ChinaStock Photo Library

Shang dynasty (16th to 11th centuries BC). Although Chinese legends speak of the Hsia dynasty, the Shang is the first whose existence is attested by archaeological and contemporary written records. Its origins are obscure. The Shang capital was reportedly moved on a number of occasions. Erh-li-t'ou, discovered near the modern city of Lo-yang in Honan province, may represent the earliest named Shang capital, Po, if not a still earlier Hsia dynasty site. A "palace" with pounded-earth foundation, fine jades, simple bronze vessels, and oracle bones have all been found there, but the question of whether this represents the Shang dynasty or its predecessor remains uncertain for lack of any written materials. At Erh-li-kang, in the Cheng-chou area in Honan province, traces have been found of a walled city that may have been the middle Shang capital referred to as Ao. Yin, the most enduring of Shang capital sites, lasting through the reigns of the last 9 (or 12) Shang kings, was located near the modern city of An-yang, in Honan province. Its discovery in 1899 by paleographers following the tracks of tomb robbers opened the way to verification of traditional accounts of the Shang dynasty and for the first scientific examination of China's early civilization. Here, recorded on oracle bones, the written documentation for the first time is rich, archival, and wide-ranging regarding activities of the theocratic Shang government. Excavations conducted near An-yang between 1928 and 1937 provided the initial training ground for modern Chinese archaeology and continued periodically after 1949. No fewer than 14 royal tombs have been unearthed near An-yang, culminating in the 1976 excavation of the first major tomb to have survived intact—that of Fu Hao, who is believed to have been a consort of the Shang king Wu-ting and a noted military leader. The Fu Hao tomb contained more than 440 bronze vessels and 590 jade objects among its numerous exquisite works.

The Fu Hao tomb

Architecture. Excavations at Lo-yang, Cheng-chou, and An-yang have revealed rammed-earth (layers of pounded earth) foundations and postholes of timber buildings with wattle and daub walls (woven rods and twigs covered and plastered with clay) and thatched roof. The largest building yet traced at An-yang is a timber hall about 30 metres (90 feet) long, the wooden pillars of which were set on stone socles, or bases, on a raised platform. Ordinary dwellings were partly sunk beneath ground level, as in Neolithic times, with deeper storage pits inside them.



Development of bronze vessel types.
From W. Fong (ed.), *The Great Bronze Age of China*, copyright © 1980 The Metropolitan Museum of Art, New York City. Alfred A. Knopf, Inc.; after a drawing by Phyllis Ward

There is no sign of the structural use of brick or stone or of tile roofs in any of the An-yang sites. Royal tombs along the banks of the Huan River to the northwest of modern An-yang consisted of huge, square, rammed-earth pits approached by two or four sloping ramps. Lined and roofed with timber, the tombs were sunk in the floor of the pit. Tomb walls and coloured impressions left on the earth by carved and painted timbers include zoomorphic motifs very similar to those on ritual bronze vessels (see below). Traces of a painted clay wall found elsewhere at An-yang, in a royal stone- and jade-carving workshop, demonstrate that buildings above ground were decorated with similar designs and indicate a uniformity of design principles and themes in all media at that time.

Sculpture. While no monumental sculpture has been found at Shang sites, Shang craftsmen carved, generally out of white marble, small, seated human figures, birds, tigers, elephants, bears, and composite creatures. Most of these pieces were made as isolated objects for display or ritual purposes, but a few are socketed, showing that they once adorned a structure. These marble sculptures are compactly modeled and decorated with volutes and squared spirals similar to those on ritual bronzes. In addition to stone carvings, a small number of pottery figurines of prisoners of war have been unearthed at An-yang, the earliest instance of the custom of putting figurines of guardians and slaves in the tomb to accompany the dead.

Pottery tomb figurines

Ritual bronzes. More than any other factor, it was the unearthing at An-yang of magnificent bronze vessels that demonstrated the power and wealth of the Shang rulers. The vessels were used in divinatory ceremonies for sacrificial offerings of meat, wine, and grain, primarily to the spirits of clan ancestors, especially those of the ruler and his family. They were probably kept in the ancestral hall of the clan, and, in some cases, they were buried with their owner.

Surprisingly, perhaps, the bronze vessels were not discussed in Shang oracle bone inscriptions. But they themselves sometimes came to bear, by late Shang times, short, cast dedicatory inscriptions, providing the name of the vessel type, of the patron, and of the ancestor to whom the vessel was dedicated. What may be a clan name is also often included, enclosed within an inscribed notched square of uncertain meaning but now called a *ya-hsing*. The common addition by early Chou times of the phrase "May sons and grandsons forever treasure and use it" provides evidence that most vessels were made originally for use in temple sacrifices rather than for burial, but other vessels, poorly cast and inscribed with posthumous ancestral names of the newly deceased, were clearly intended for the tomb.

The right to cast or possess these vessels was probably originally confined to the royal house itself but later was bestowed upon local governors set up by the ruler; still later, in the Chou dynasty, the right was claimed by rulers of the feudal states and indeed by anyone who was rich and powerful enough to cast his own vessels.

Bronze vessel types

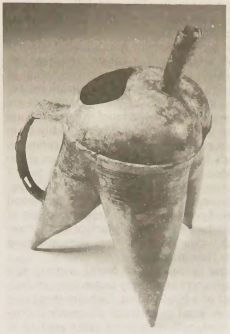
The vessel types are known today either by names given them in Shang or Chou times that can be identified in contemporary inscriptions, such as the *li, ting,* and *hsien,* or by names, such as *yu, chia,* and *kuang,* given them by later Chinese scholars and antiquarians. The vessels may be grouped according to their presumed function in sacrificial rites. For cooking food the main types are the *li,* a round-bodied vessel with a trilobed base extending into three hollow legs; its cousins, the *ting,* a hemispheric vessel on three solid legs, and *fang-ting,* a square vessel standing on four legs; and the *hsien,* or *yen,* a steamer consisting of a bowl placed above a *li* tripod, with a perforated grate between the two. For offering food, the principal vessel was the *kuei,* a bowl placed on a ring-shaped foot, like a modern-day wok. The word *tsun* embraces wine containers of a variety of shapes. Among vessels for heating or offering wine are the *yu,* a covered bucket with swing handle; the *chia,* a round tripod or square quadruped with a handle on the side and raised posts with caps rising from its rim; the related *chüeh,* a smaller beaker on three legs, with an extended pouring spout in front, a pointed tail in the rear, a side handle, and posts with caps; the *ho,* distinguished by its cylindrical pouring spout; the *kuang,* resembling a covered gravy boat; and the elegant trumpet-mouthed *ku.* Vessels for ablutions include the *p'an,* a large, shallow bowl. The shapes of the round-bodied vessels were often derived from earlier pottery forms; the square-section vessels, with flat sides generally richly decorated, are thought to derive from boxes, baskets, or containers of carved wood or bone.

Other objects connected with the rites were bronze drums and bells. Weapons and fittings for chariots, harness, and



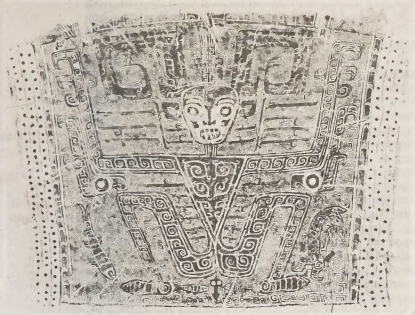
T'ao-t'ieh mask from a first-phase *ting*, 13th–10th centuries BC.

other utilitarian purposes also were made of bronze. Bronze vessels were cast not by the lost-wax process (using a wax mold), as formerly supposed, but in sectional molds, quantities of which have been found at Shang sites. In this complex process, which reflects the Chinese early mastery of the ceramic medium, a clay model of the body is built around a solid core representing the vessel's interior; clay molding is used to encase the model, then sliced into sections and removed; the model is eliminated; the mold pieces are reconstructed around the core, using metal spacers to separate mold and core; molten bronze is poured into the hollow space. Legs, handles, and appended sculpture are often cast separately and integrated in a later, lock-on pour. Surface decoration may be added to the model surface before the mold is applied, requiring a double-transfer from clay to clay to metal, or added in reverse to the mold surface after its removal from the model, with an incised design on the mold yielding a raised design on the metal surface. Ritual vessels range from about 15 centimetres to over 130 centimetres in height with weights up to 875 kilograms (1,925 pounds). The intricacy and sharpness of the decoration shows that by the end of the 2nd millennium BC the art of bronze casting in China was the most advanced in the world.



Bronze *he* (Style I), Shang dynasty. In the Asian Art Museum of San Francisco. Height 22.9 cm.
Asian Art Museum of San Francisco, The Avery Brundage Collection (B60B53)

Sen-oku Hakko kan, Kyoto



"Shaman figure" cast on a ceremonial bronze drum (Style IV) rubbing, c. 12th century BC, Shang dynasty. In the Sen-oku Hakko kan, Sumitomo Collection, Kyōto, Japan.

While many Shang ritual bronzes are plain or only partly ornamented, others are richly decorated with a variety of geometric and zoomorphic motifs, and a small number take the form of a bird or animal. The dominating motif is the *t'ao-t'ieh,* seen either as two stylized creatures juxtaposed face-to-face or as a single creature with its body splayed out on both sides of a masklike head. The term *t'ao-t'ieh* first appeared in the late Chou and is perhaps related to eclipse mythology and the idea of renewal. Sung dynasty antiquarians offered the unlikely interpretation that it represented a warning against gluttony. Alternative modern suggestions are that it was a fertility symbol like the later Chinese dragon, bestowing longevity on the ruling clan; that it was a fierce spirit which protected the rites and the participants from harm; that it embodied a variety of creatures related to the ceremonial sacrifices; that it was totemic or related to shamanic empowerment; or that its dual structure represented the inseparable forces of creation and destruction. Other creatures on the bronzes are the *k'uei* (each like half of the doubled *t'ao-t'ieh*), tiger, cicada, snake, owl, ram, and ox. In later times, the tiger represented nature's power, the cicada and snake symbolized regeneration, the owl was a carrier of the soul, and the ram and ox were chief animals of ancestral sacrifices. It is not known whether these meanings were attached to the creatures on Shang bronzes, for no Shang writing addresses the issue, but it seems likely that they had a more than purely decorative purpose. There is no suggested environmental setting for these creatures. The human figure appears only rarely in Shang bronzes, usually in the grasp of these powerful zoomorphic creatures.

The *t'ao-t'ieh*

Evolution of the Shang style

The art of the Shang bronzes evolved from technically simple, albeit sometimes quite elegant, thinly cast vessels, clearly revealing ceramic prototypes. It reached a climax of sculpturesque monumentality at the end of the dynasty, reflecting a long period of peace and stability at An-yang. In the early 1950s the scholar Max Loehr identified five phases or styles in the evolution of Shang bronze surface decor and casting techniques. The thin-walled vessels of Style I typically carry a narrow register of zoomorphic motifs that are more abstract in appearance than motifs of later times; the motifs are composed of thin, raised lines created by incision on the production molds. Style II zoomorphic forms are composed of broad, flat bands in narrow horizontal registers, incised on the model, often on a raised band of ceramic appliqué. In Style III, dense curvilinear designs derived from those of the previous phase begin to cover much of the surface of an increasingly thick-walled vessel, and the zoomorph becomes increasingly difficult to discern. The main zoomorphic motifs of Style IV, although flush to the surface of the vessel (ex-

The Metropolitan Museum of Art, New York City, on loan from the People's Republic of China; photography by Seth Joel



Ceremonial bronze *lei* (Style III), c. 13th century BC, Shang dynasty. In the Palace Museum, Peking. Height 52 cm.

clusive of appended heads, handles, and fully sculptural attachments), become clearly distinguishable as set against a dense spiral background known as "thunder pattern" (*lei-wen*); in this phase, with similar spirals placed sparsely over the zoomorph, which itself is constructed from the same linear vocabulary, an intricate decorative system of interactive forms, rich in philosophical implications, begins to reach maturity. In Style V, the main motifs are set forth in increasingly bold plastic relief through the use of ceramic appliqué upon the model. Style I bodily form clearly reveals conceptualization derived from ceramics, while Style V vessels fully utilize the sculpturesque possibilities of the molded-bronze technology. Styles I and II appear at Cheng-chou; Style III appears at both Cheng-chou and early An-yang; and Styles IV and V are found in the An-yang period only. Pre-Style I vessels, ceramic in form, thin-walled, and with little or no surface decor, have been found at Erh-li-t'ou near Lo-yang, demonstrating early Shang or even Hsia origins.

Development of pottery glazes

Ceramics, jade, and lacquer. The Shang dynasty saw several important advances in pottery technology, including the development of a hard-bodied, high-fire stoneware and pottery glazes. A small quantity of stoneware is covered with a thin, hard, yellowish green glaze applied in liquid form to the vessel. Shang potters also developed a fine soft-bodied white ware, employing kaolin (later used in porcelain); this ware was probably for ceremonial use and was decorated with motifs similar to those on the ritual bronzes. Much cruder imitations of bronze vessels also occur in the ubiquitous gray pottery of the Shang dynasty.

In the Shang dynasty, and particularly at An-yang, the craft of jade carving made a notable advance. Ceremonial weapons and fittings for bronze weapons were carved from jade; ritual jades included the *pi, ts'ung,* and symbols of rank. Plaques and dress ornaments were carved from thin slabs of jade, but there are also small figurines, masks, and birds and animals carved in the round, some of these perhaps representing the earliest examples of *ming-ch'i* ("spirit vessels"), artistic figures substituted for live victims buried in order to serve the deceased.

Coffins, chariots, furniture, and other objects found in Shang tombs were often lacquered, and lacquer was used to fix inlays of shell and coloured stone. Small decorative and functional objects such as hairpins, finials, buttons, and knife handles were often fashioned of bone or ivory, sometimes inlaid with chips of turquoise.

Chou dynasty (11th century–255 BC). The arts of the Chou dynasty, the longest in Chinese history, reflect the profound changes that came over Chinese society during nearly 800 years. The first Chou rulers took over the Shang culture to the extent that the earliest bronze vessels bearing Chou inscriptions might, from their style, have been made in the Shang dynasty. The Chou kings parceled out their expanding territory among feudal lords, each of whom was free to make ritual objects for his own court use. As the feudal states rose in power and independence, so did the central Chou itself shrink, to be further weakened by the eastward shift of the capital from sites in the Wei River valley near modern-day Sian to Lo-yang in 771 BC. Thereafter, as the Chou empire was broken up among rival states, many local styles in the arts developed. The last three centuries of the Chou dynasty, known as the Warring States period, saw a flowering of the arts in many areas. The breakdown of the feudal hegemony, the growth of trade between the states, and the rise of a rich landowning and merchant class all brought into existence new patrons and new attitudes that had a great influence on the arts and crafts.

Architecture. Remains of a number of Chou cities have been discovered, among them capitals of the feudal states. They were irregular in shape and surrounded by walls of rammed earth. Some long defensive walls also have been located, the largest being one that protected the state of Ch'i from Lu to the south, stretching for more than 500 kilometres (300 miles) from the Huang Ho to the sea. Ch'u had a similar wall along its northern frontier.

Foundations of a number of palace buildings have been found in the cities, including, at Hui-hsien, the remains of a hall 26 metres (85 feet) square, which was used for ancestral rites in connection with an adjacent tomb—an arrangement that became common in the Han dynasty. An important late Chou structure used for the conduct of state rituals was the Ming-t'ang ("Spirit Hall"), discussed in Chou literature but not yet known through excavations. Late Chou texts also describe platforms or towers, *t'ai,* made of rammed earth and timber and used as watchtowers, as treasuries, or for ritual sacrifices and feasts, while pictures engraved or inlaid on late Chou bronze vessels show two-story buildings used for this type of ritual activity. Some of these multistory buildings are now understood, through modern excavations of two- and three-story Ch'in and Han palaces and of state ritual halls at Hsien-yang, Sian, and Lo-yang, to have been constructed around a large, raised pounded-earth core that structurally supported upper building levels and galleries and into which lower-level chambers were inserted.

Multistory buildings

The origins of the Chinese bracketing system also are found on pictorial bronzes, showing a spreading block (*tou*) placed upon a column to support the beam above more broadly, and in depictions of curved arms (*kung*) attached near the top of the columns, parallel to the building wall, extending outward and up to help support the beam; however, the block and arms were not yet combined to create traditional Chinese brackets (*tou-kung*) or to achieve extension forward from the wall. Roof tiles replaced thatch before the end of the Western Chou (771 BC), and bricks have been found from early in the Eastern Chou.

Ritual bronzes and related works. The ritual bronzes of the early Hsi (Western) Chou continued the late An-yang tradition; many were made by the same craftsmen and by their descendants. Even in the predynastic Chou period, however, new creatures had appeared on the bronzes, notably a flamboyant long-tailed bird that may have had totemic meaning for the Chou rulers, and flanges had begun to be large and spiky. By the end of the 9th century, moreover, certain Shang shapes such as the *chüeh, ku,* and *kuang* were no longer being made, and the *t'ao-t'ieh* and other Shang zoomorphs had been broken up and then dissolved into volutes or undulating meander patterns encircling the entire vessel, scales, and fluting, with little apparent symbolic intent.

From the outset of Chou rule, vessels increasingly came to serve as vehicles for inscriptions that were cast to record events and report them to ancestral spirits. An outstanding example, excavated near Sian in 1976, was dedicated by a Chou official who apparently had divined the date for the successful assault upon the Shang and later used his reward money to have the bronze vessel cast. By late Chou times, a long inscription might have well over 400 characters.

Vessel shapes, meanwhile, had become aggressive or heavy and sagging, and the quality of the casting is seldom as high as in the late Shang. These changes, completed by the 8th century BC, mark the middle Chou phase of bronze design.

Bronzes of the Eastern Chou period

The bronzes of the Tung (Eastern) Chou period, after 771 BC, show signs of a gradual renaissance in the craft and much regional variation, which appears ever more complex as more Eastern Chou sites are unearthed. Vessels from Hsin-cheng in Honan (8th to 6th century BC) reveal a further change to more elegant forms, often decorated with an allover pattern of tightly interlaced serpents. The aesthetic tendency toward elaboration was given further stimulus by the introduction of the lost-wax method of production (by the late 7th century BC), leading quickly to zealous experiments in openwork design that are impressive technically though heavy in appearance and gaudy in effect. The style of bronzes found at Li-yü in Shansi (*c.* 6th–5th century BC) is much simpler, more compact, and unified; the interlaced and spiral decoration is flush with the surface. Thereafter, until the end of the dynasty, the bronze style became increasingly refined; the decoration was confined within a simpler contour, the interlacing of the Hsin-cheng style giving way to the fine, hooked "comma pattern" of the vessels of the 5th and 4th centuries BC. By this time, bronze decor had come under the influence of textile patterns and technique, particularly embroidery, as well as of lacquer decor, suggesting

the decline from primacy of the bronze medium. Bronzes thus decorated have been found chiefly in the Huai River valley.

Bronze bells are exemplified by an orchestral set of 64 bells, probably produced in Ch'u and unearthed in 1978 from a royal tomb of the Tseng state, at Lei-ku-tun near Sui-hsien in Hupeh province. The bells were mounted on wooden racks supported by bronze human figurines. They are graded in size (from about 20 to 150 centimetres in height) and tone (covering five octaves), and each is capable of producing two unrelated tones according to where it is struck. Gold-inlaid inscriptions on each bell present valuable information regarding early musical terms and performance, while a 65th bell is dedicated by inscription from the king of Ch'u to Marquis I of Tseng, the deceased, and bears a date equivalent to 433 BC.

Finally, in vessels from the rich finds at Chin-ts'un near Lo-yang, all excrescences are shorn away; the shapes have a classic purity and restraint, and the decoration consists of geometric patterns of diagonal bands and volutes. The taste of the new leisured class is shown in objects that were not merely useful but finely fashioned and beautiful in themselves: ritual and domestic vessels, weapons, chariot and furniture fittings, ceremonial staff ends, bracelets, and the backs of mirrors. Monster masks attaching ring handles are reminiscent of the Shang *t'ao-t'ieh*, the first sign of a deliberate archaism that from time to time thenceforward gave a special flavour to Chinese decorative art.

The wealth and sophistication of late Chou culture is shown by exquisite craftsmanship, while the new techniques of cast openwork and many of the works executed with inlays of gold, silver, jade, glass, and semiprecious stones also indicate the increasing commercial interaction and artistic fascination of the Chinese with the tribal peoples to their north. Bronze garment hooks worn at the shoulder were often fashioned in the form of animals, reflecting the artistic style of China's nomadic neighbours, who through the Eastern Chou and Han dynasties exerted pressure on its northern frontiers and who both influenced and were influenced by Chinese culture in this period.

Bronze objects

Bronze mirrors were used in ancient China not only for toiletry but also as funerary objects, in accordance with the belief that a mirror was itself a source of light and could illuminate the eternal darkness of the tomb. A mirror also was thought of as a symbolic aid to self-knowledge. Chinese mirrors are bronze disks polished on the face and decorated on the back, with a central loop handle or pierced boss to hold a tassel. The early ones were small and worn at the girdle; later they became larger and were often set on a stand. A bronze disk found in a tomb at An-yang may have been a mirror. There is less doubt about the small disks from an 8th-century-BC tomb at Shang-ts'un-ling in Honan, believed to be the earliest mirrors yet found in China. Mirrors, however, were not widely used until the 4th and 3rd centuries BC. Shou-chou, in the state of Ch'u, was a centre for the manufacture of late Chou mirrors.

Sculpture. Chou sculpture, like that of the Shang, is typically small in scale and occurs most frequently in the medium of sculpturesque ritual bronze vessels. Examples from the middle Chou period are usually stiffly formal, but the forceful spatial rendering and emergent naturalism characteristic of Ch'in and early Han dynasty sculpture and seen in a massive late 3rd-century-BC inlaid bronze rhinoceros from Hsing-p'ing, Shensi province, had already appeared by the Warring States era. Inlaid bronze sculpture from the northeastern state of Chung-shan (late 4th century BC) reveals the influence of neighbouring nomadic art, and, if not as fully naturalistic, these works are nonetheless remarkably dynamic and alert, one notable example depicting a tiger capturing a hapless deer. No less fierce and even more bizarre are some of the lacquered sculptures from the southern state of Ch'u, such as the monstrous creature with real deer horns, bulging eyes, and long tongue devouring a snake, from about the 3rd century BC, excavated at Hsin-yang in Honan province. Simple pottery grave figurines also have been found at late Chou sites in northern China; and attenuated, formalized wooden figurines of servants and attendants, with details of dress painted on them, have been found in graves of the state of Ch'u.

Painting and pictorial arts. Practically nothing survives of Chou painting, although from literary evidence it seems that the art developed considerably, particularly during the period of the Warring States. Palaces and ancestral halls were decorated with wall paintings. The most significant development of the late Chou, and among the most revolutionary of all moments in Chinese art, was the emergence of a representational art form, departing from the ritualized depiction of fanciful and usually isolated creatures of the Shang and early to middle Chou. In decorating ceremonial objects, artists began to depict the ceremonies themselves, such as ancestral offerings in temple settings, as well as ritual archery contests (important in the recruitment and promotion of officials), agriculture and sericulture, hunting, and the waging of war—all activities vital to a well-ordered state. Such representations were cast with gold or silver inlay or engraved onto the sides of bronze vessels, most notably the *hu*, where all these themes might be combined on a single vessel. This conceptual transformation began by the late 6th century BC, at about the same time that Confucius and other philosophers initiated humane speculation on the nature of statecraft and social welfare.

Emergence of representational art

The early representation of landscape, indicated only crudely on bronzes, appears in more sophisticated fashion on embroidered textiles of the 4th–3rd centuries BC from such south-central Chinese sites as Ma-shan, near Chiangling in the state of Ch'u (modern Hupeh province). There, as in Han dynasty art to follow, landscape is suggested by rhythmic lines, which serve as mountain contours to spatially organize a variety of wild animals in front and back and which, while structurally simple, convey in linear fashion a sophisticated concept of mountain landscape as fluid, dynamic, and spiritual.

Some of these motifs and, perhaps, the early treatment of landscape itself may derive in both theme and style from foreign sources, particularly China's northern nomadic neighbours. Those scenes concerned with ceremonial archery and ritual offerings in architectural settings, sericulture, warfare, and domestic hunting, however, seem to be essentially Chinese. These renditions generally occur with figures in flattened silhouette, spread two-dimensionally and evenly over most of the available pictorial surface. But, by the very late Chou, occasional examples, such as the depiction of a mounted warrior contending with a tiger, executed in inlaid gold and silver on a bronze mirror

Wang Lu/ChinaStock Photo Library



Drawing of ancestral offering scenes (ritual archery, sericulture, hunting, and warfare) cast on a ceremonial bronze *hu*, 6th–5th century BC, Chou dynasty. In the Palace Museum, Peking.

from Chin-ts'un (*c.* 3rd century BC, Hosokawa collection, Tokyo), suggest the emerging ability of artists to conceive of two-dimensional images in terms of implied bulk and spatial context.

Early paintings on silk

The few surviving Chou period paintings on silk—from about the 3rd century BC, the oldest in all East Asia—were produced in the state of Ch'u and unearthed from tombs near Ch'ang-sha. One depicts a woman, perhaps a shamaness or possibly the deceased, with a dragon and phoenix; one depicts a gentleman conveyed in what appears to be a dragon-shaped boat; and a third, reported to be from the same tomb as the latter, is a kind of religious almanac (the earliest known example of Chinese writing on silk) decorated around its border with depictions of deities and sacred plants.

Jade. In the Chou, production of jade *pi, ts'ung,* and other Shang ritual forms was continued and their use systematized. Differently shaped sceptres were used for the ranks of the nobility and as authority for mobilizing troops, settling disputes, declaring peace, and so on. At burial, the seven orifices of the body were sealed with jade plugs and plaques. Stylistically, Chou dynasty jades first continued Shang traditions, but then, just as the bronzes did, they turned toward looser, less systematic designs by middle Chou times, with zoomorphic decor transformed into abstract meander patterns. This breakdown of formal structure continued to the end of the dynasty.

The introduction of iron tools and harder abrasives in the Eastern Chou led to a new freedom in carving in the round. Ornamental jades, chiefly in the form of sword and scabbard fittings, pendants, and adornment for the clothing, were fashioned into a great variety of animals and birds, chiefly from flat plaques no more than a few millimetres thick.

Glass. Glass was already in use in China in the Western Chou period, attested by beads found in tombs in Sian and Lo-yang of the 9th and 8th centuries BC. In the Eastern Chou it was sometimes used as a cheap substitute for jade *pi* disks and sword fittings and as an inlay in garment hooks and various ornamental objects. While some glass beads were certainly imported from western Asia, the craft of glassmaking may have begun in China independently as a by-product of the manufacture of pottery glaze.

Ceramics. Early Western Chou pottery, like the bronzes, continued the Shang tradition at a somewhat lower technical level, and the soft white Shang pottery disappeared.

The Metropolitan Museum of Art, New York City, on loan from the People's Republic of China; photograph by Seth Joel



Kneeling archer from the tomb of Emperor Shih Huang-ti, Lin-t'ung, near Sian, Shensi province, *c.* 210 BC, Ch'in dynasty. Terracotta with traces of paint. In the Museum of Ch'in Dynasty Terracotta Warriors and Horses, Lin-t'ung. Height 1.2 m.

Stemmed offering dishes, *tou,* were made in a hard stoneware dipped or brushed over with a glaze ranging from gray to brownish green. The fact that some of the richest finds of high-fired glazed wares have been made not in Honan but at Yi-ch'i in Anhwei shows that the centre of advance in pottery technology was beginning to move, with the growth of population, to the lower Huai and Yangtze valleys. Crude attempts also were made to give pottery the appearance of bejeweled metal by covering *tou* stands with lacquer inlaid with shell disks.

In the second half of the dynasty the range of pottery types and techniques was greatly extended. A low-fired pottery was produced in Honan primarily for burial. Some of it is white, and some is covered with slip, or liquid clay, and painted, reviving an ancient northern China tradition. At Hui-hsien has been found a soft-bodied, black burnished ware, sometimes decorated with scrolls and geometric motifs scratched through the polished surface. In the period of the Warring States, a soft earthenware covered with green lead glaze was made in northern China for burial. In the lower Yangtze valley an almost porcelaneous stoneware was developing, covered with a thin feldspathic glaze, the ancestor of the celadon glaze of the T'ang dynasty and later. Another technique, which appears in the glazed wares of Chekiang and Kiangsu and was to persist in the southern pottery tradition for many centuries, was the stamping of regular, repeated motifs over the surface of the vessel before firing.

Development of an almost porcelaneous stoneware

Ch'in (221–206 BC) and Han (206 BC–AD 220) dynasties. In 221 BC the ruler of the feudal Ch'in state united all China under himself as Ch'in Shih Huang-ti ("First Sovereign Emperor of Ch'in") and laid the foundation for the long stability and prosperity of the succeeding Han dynasty. His material accomplishments were the product of rare organizational genius, including centralizing the Chinese state and its legal system, unifying the Chinese writing script and its system of weights and measures, and consolidating many of the walls of northern China into an architectural network of beacon towers able to spot any suspicious military movement and relay messages across the territory in a single day. However, his means were brutal and exhausted the people, and the dynasty failed to survive his early death.

The Hsi (Western) Han (206 BC–AD 25), with its capital at Ch'ang-an (near modern Sian), reached a climax of expansive power under Wu-ti (ruled 141/40–87/86 BC), who established colonies in Korea and Indochina and sent expeditions into Central Asia, which both made Chinese arts and crafts known abroad and opened up China itself to foreign ideas and artistic influences. After the period of the usurping Hsin dynasty (AD 9 to 25), the Tung (Eastern) Han, with its capital at Lo-yang, recovered something of the dynasty's former prosperity but was increasingly beset by natural disasters and rebellions that eventually brought about its downfall. The art of the Han dynasty is remarkable for its variety and vigour, the product of foreign contacts, of a national unity in which many local traditions flourished, and of the patronage of a powerful court and the new, wealthy landowning and official classes.

Architecture. While little remains except walls and tombs, much can be learned about Han architecture from historical writings and long descriptive poems, *fu.* This was an era of great palace buildings. The first Ch'in emperor undertook the building of a vast palace, the A-fang or O-pang, whose main hall was intended to accommodate 10,000 guests in its upper story. He also copied the palaces and pavilions of each of the feudal lords he had defeated; these buildings stretched more than seven miles along the Wei River and were filled with local lords and women captured from the different states.

The first emperor's tomb was part of a city of the dead that covered nearly 2 square kilometres (0.75 square mile) and was surrounded by double walls, with numerous gates, corner towers, and a ceremonial palace. The mausoleum itself was surmounted by an artificial mound, a feature not known in the Shang or early Chou and first found among the 4th–3rd-century-BC tombs near Chiang-ling in Hupeh province. About 43 metres high, this tumulus was shaped like a triple-layered truncated pyramid symbolizing

heaven, man, and earth. The tomb, which has not yet been excavated, reportedly featured a large chart of the heavens painted on its domed vault and a three-dimensional representation of the earth below, with rivers of liquid mercury driven by mechanical contrivances. Excavations around the tomb have uncovered a large protective "spirit army" of some 7,000 life-size terra-cotta figurines, along with 400 horses and 100 chariots, placed in battle formation in a series of pits beneath the nearby fields. Molded in separate sections, assembled, then fully painted, these warrior figures are executed in minute and realistic detail and provide evidence of an early naturalistic sculptural tradition scarcely imagined before their discovery in 1974. For the heads, up to 30 different models were used, and each was hand-finished to give further variety. Excavated later, in 1982, was a pair of precisely engineered bronze replicas of the Imperial chariot (104 centimetres high, with considerable gold and silver inlay), each with charioteer and four horses, possibly indicating the presence of a still larger underground stable of such figures.

Terra-cotta "spirit army"

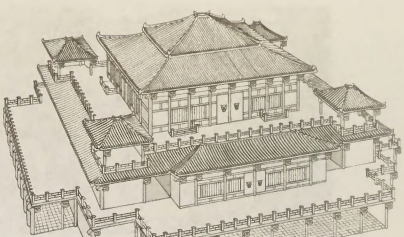
The main audience hall of the Western Han Wei-yang palace was said to have been about 120 metres long by 35 metres deep, possibly smaller than its largest Ch'in predecessor yet much larger than its equivalents in the Peking palace today. From the Chou dynasty through the Yüan, no architectural structure called forth more intense consideration than the so-called Ming-t'ang ("Spirit Hall"), the predecessor of Peking's Temple of Heaven. The site of the Han ritual hall, in the southern suburbs of Han dynasty Ch'ang-an, was excavated in 1956–57. Translating traditional ritual values into symbolic architecture, the Ming-t'ang was surrounded by an outer circular moat and set on a circular foundation (the two circles together forming a disk, or *pi*, symbolic of heaven) and was further enclosed within an intermediate rectilinear colonnade (symbolic of earth). The three-story hall itself (the number three signifying heaven, man, and earth) was built around a raised earthen core. It is thought to have been a composite ritual structure that included a royal academy on the first floor; a second floor divided into nine zones, corresponding to the four seasons and the "five phases" theory of change, with five inner shrines and with outer spaces for monthly ritual offerings; and a third-floor central hall surrounded by a terrace (*ling-t'ai*, or "spirit platform") for observation of the heavens and regulation of the calendar.

The Han palaces were set about with tall timber towers (*lou*) and brick or stone towers (*t'ai*) used for a variety of purposes, including the display and storage of works of art. Ceramic representations of Han architecture provide the first direct evidence of true bracketing, with simple brackets projecting a single step forward (and sometimes several steps upward) from the wall in order to support the roof projection.

Han tombs

Han tombs are among the most elaborate ever constructed in China. In some localities they are of timber, but more often they are of brick or stone, divided into several chambers, and covered with a corbeled vault or more rarely a true arched vault. The tombs of the Han emperors were enclosed in gigantic earth mounds that are still visible today, but some royal tombs began the later practice of burial in hollowed-out natural hills. Many Han tombs were decorated with wall paintings, with more permanent and expensive stone reliefs, or with stamped or molded bricks.

The most remarkable excavated tomb of the period belonged to the wife of a mid-level aristocrat, one of three family tombs of the governor of Ch'ang-sha found in a suburb of that southern city, Ma-wang-tui, and dating from 168 BC or shortly after. Small in scale but richly equipped and perfectly preserved, the wooden tomb consists of several outer compartments for grave goods tightly arranged around a set of four nested lacquered coffins. An outer layer of sticky white kaolin clay prevented moisture from penetrating the tomb, and an inner layer of charcoal fixed all the available oxygen within a day of burial, so the deceased (Hsin Chui, or Lady Tai, the governor's wife) was found in a near-perfect state of preservation. Included among the grave goods, which came with a written inventory providing contemporaneous terminology, are the finest caches yet discovered of early Chinese silks (gauzes and damasks, twills and embroideries, including many whole garments) and lacquerwares (including wood-, bamboo-, and cloth-cored examples), together with a remarkable painted banner that might have been carried by the shaman in the funerary procession (see below).



Reconstructive elevation of the Ming-t'ang state ritual hall, Ch'ang-an (Sian), Shensi province, *c.* 1st century BC, Han dynasty.
From N.S. Steinhardt *et al.*, *Chinese Traditional Architecture* (1984); © China Institute in America, Inc., courtesy of China Institute Gallery, New York City

Painting and related arts. Literature and poetry indicate that the walls of palaces, mansions, and ancestral halls were plastered and painted. Themes included figure subjects, portraits, and scenes from history that had an ethical or didactic purpose. Equally popular were themes taken from folk and nature cults that expressed the beliefs of popular Taoism. The names of the painters are generally not known. Artists were ranked according to their education and ability from the humble craftsmen painters (*hua-kung*) up to the painters-in-attendance (*tai-chao*), who had high official status and were close to the throne, a bureaucratic system that lasted into the Ch'ing dynasty.

In addition to wall paintings, artists painted on standing screens, used as room dividers and set behind important personages, and on long rolls of silk. Paper was invented in the Han dynasty, but it is doubtful whether it was much used for painting before the 3rd or 4th century AD.

Han tomb paintings

The most important painted tombs have been found at Lo-yang, where some are decorated with the oldest surviving historical narratives (1st century BC); at Wang-tu in Hopeh (Eastern Han), where they are adorned with figures of civil and military officials; and at Liao-yang in Liaoning, where the themes include a feasting scene, musicians, jugglers, chariots, and horsemen. The celebrated bricks taken from a tomb shrine of the Eastern Han (now in the Museum of Fine Arts, Boston) depict gentlemen in animated conversation, elegant and individualized and rendered with a sensitive freedom of movement.

Funerary slabs also reflect the variety of Han pictorial art. The most famous are those from tomb shrines of the Wu family at Chia-hsiang in Shantung, dated between about AD 147 and 168. The subjects range from the attempted assassination of the first Ch'in emperor to feasting and mythological themes. Although they are depicted chiefly in silhouette with little interior drawing, the effect is lively and dramatic. These well-known works have been generally taken as representative of Han painting style since their discovery in 1786. They are now understood, however, to be very conservative in style, even archaistic, perhaps with the intent of advertising the sponsoring family's chaste attachment to the pure and simple virtues of past times. A far earlier painting, a funerary banner from about 168 BC, excavated in 1972 at Ma-wang-tui, reveals how much more sophisticated early Han and even late Chou painting must have been. Painted with bright, evenly applied mineral pigments and fine, elegant brush lines on silk, the banner represents a kind of cosmic array, with separate scenes of a funerary ceremony, the underworld, and the ascent of the deceased (the Lady Tai mentioned



Rubbing of a stone relief from the tomb shrine (*tz'u*) of Wu Liang, Shantung province, Eastern Han dynasty (AD 25–220). In the Department of Art and Archaeology, Princeton University, New Jersey. U.S. 70 cm × 146 cm.

By courtesy of the Department of Art & Archaeology, Princeton University, New Jersey

above) to a heavenly setting filled with mythic figures. It contains stylistic features not previously seen before the 4th century AD, creating spatial illusion through foreshortening, overlapping, and placement upon an implied ground plane, as well as suggesting certain lighting effects through contrasting and modulated colours.

Han landscape painting

Han landscape painting is well represented by the lacquer coffins of Lady Tai at Ma-wang-tui, two of which are painted with scenes of mountains, clouds, and a variety of full-bodied human and animal figures. Two approaches are used: one, more architectonic, uses overlapping pyramidal patterns that derive from the bronze decor of the late Chou period; the other continues the dynamic linear convention already noted on the embroidered textiles from Chiang-ling, in the Warring States period, as well as on late Chou painted lacquers, on inlaid bronze tubes used as canopy fittings for chariots, and on woven silks found at Noin-ula, in Mongolia. Elsewhere, in the late Han, a new feeling for pictorial space in a more open outdoor setting appeared on molded bricks decorating tombs near Ch'eng-tu; these portrayed hunting and harvesting, the local salt-mining industry, and other subjects.

Lacquer. By the Han dynasty, lacquer production was chiefly carried on at Ch'ang-sha and in four regional factories in Shu (modern Szechwan) under government control. In addition to the fine lacquerwares excavated from tombs in Ch'ang-sha, splendid products of the Szechwan workshops, bearing inscriptions dated between 85 BC and AD 71, have been found in tombs of Chinese colonists at Lo-lang (Nangnang) in North Korea, and pieces of Han lacquerware have been found as far afield as northern Mongolia and Afghanistan.

The different stages of Han lacquer manufacture were divided among a number of specialized craftsmen. The *su-kung,* for example, prepared the base, which might be of hemp cloth, wood, or bamboo basketwork; after priming, the base was covered with successive layers of lacquer by the *hsiu-kung.* The top layer, applied by the *shang-kung,* was polished and so prepared for the painter, *hua-kung,* who decorated it. Others might inlay the design or engrave through the top coating to another colour beneath it, add gilding, and write or engrave an inscription. A wine cup found at Lo-lang bears an inscription giving its capacity, the names of the people concerned in its manufacture, a date equivalent to AD 4, and place of origin, the "Western Factory" in Shu Commandery.

Among the most celebrated examples of Han lacquer painting is a basket found at Lo-lang (National Museum, Seoul), decorated with 94 small figures of paragons of filial piety, virtuous and wicked rulers, and ancient worthies.

Zhang Ping/ChinaStock Photo Library



Drawing of a landscape scene from a bronze chariot canopy fitting, from Ting-hsien, Hopeh province, *c.* 2nd–1st century BC, Western Han dynasty. In the Hopeh Provincial Museum, Wu-han. Height 26.5 cm.

Development of silk weaving into a major industry

Textiles. Silk weaving became a major industry and one of China's chief exports in the Han dynasty. The caravan route across Central Asia, known as the Silk Road, took Chinese silk to Syria and on to Rome. In the 4th century BC, the Greek philosopher Aristotle mentions sericulture on the island of Cos (Kos), but the art was evidently lost and reintroduced into Byzantium from China in the 6th century AD. Chinese textiles of Han date have been found in Egypt, in graves in northern Mongolia (Noin-ula), and at Lou-lan in Chinese Turkistan. Silk was used by Han rulers as diplomatic gifts and to buy off the threatening nomads, as well as to weaken them by giving them a taste of luxury.

Early Han textiles recovered from Ma-wang-tui show the further development of the weaving traditions already present at Ma-shan in the late Chou, including brocade and embroidery, gauze, plain weaves, and damasks. Later finds elsewhere, however, are limited chiefly to damasks, very finely woven in several colours with patterns that generally repeat about every five centimetres. These designs are either geometric, the zigzag lozenge being the most common, or consist of cloud or mountain scrolls interspersed with fabulous creatures and sometimes with auspicious characters. The rectilinear patterns were transmitted from woven materials to Lo-yang bronze mirrors and appeared in paintings on both lacquer and silk; and the curvilinear scroll patterns, which are not natural to weaving, were probably adapted for embroidery from the rhythmic conventions of lacquer painting, which also provided scroll motifs for inlaid bronzes and paintings on silk. Thus, there was an interaction among the various media of Han dynasty arts that accounts for their unity of style.

Sculpture. The Han dynasty saw the creation of the first major works of stone tomb sculpture in China. It is possible that the idea of a processional way leading to the tomb, lined with monumental stone carvings of animals and guardian figures, came to China through its new contacts with western and Central Asia. Thought to be the earliest such figures are a group from the tomb of Huo Ch'ü-ping, buried in 117 BC next to the tomb of Emperor Wu-ti, in whose name he won many battles of conquest in Central Asia. The best-known of these sculptures is a crude and static yet monumentally impressive horse standing over a trampled barbarian. The Chinese tomb sculptures that followed this early example never attained much more in the way of dynamic expression. Yet Han artists were capable of a far more mobile conception, as seen in the famous bronze "flying horse" from a 2nd-century-AD military tomb at Lei-t'ai near Wu-wei, Kansu province. This work represents the lively tradition of figural bronze and ceramic sculpture unearthed in vast quantities from Han tombs. Called *ming-ch'i* ("spirit vessels"), these sculptures were meant to function as the real objects they depict in the service of the deceased. The ceramic *ming-ch'i* include servants, guards, soldiers, dancers, jugglers, and acrobats, frequently modeled with all the vitality of a popular art; they give a vivid picture of daily life in the Han dynasty and represent Chinese sculpture at its very best. There also are models of farmhouses and multistoried towers, pigpens and duck ponds, cooking stoves, and a wide variety of food and wine vessels. Ceramic *ming-ch'i* are generally of low-fired earthenware, often covered with chalk or a white slip and then painted; the painted food and wine vessels often imitate both the shape and the scrolled decoration of inlaid bronze or painted lacquer vessels.

Ritual bronzes and mirrors. Already by late Chou times, the more expensive medium of lacquer was often used in place of bronze. Nevertheless, some bronze vessels were still made for sacrificial rites, and other bronze objects such as lamps and incense burners also were made for household use. The "hill censer" (*po-shan hsiang-lu*) was designed as a miniature, three-dimensional mountain of the immortals, usually replete with scenes of mythic combat between man and beasts suggesting the powerful forces of nature that only the Taoist adept could tame. Sacred vapours emanating from materials burned within were released through perforations in the lid (hidden behind the mountain peaks). Cosmic waters were depicted lapping at the base of the hills, conveying the sense of an island; and the whole was set on a narrow stem that thrust the mountain upward as if it were an axis of the universe. Such censers might have been used in ceremonial exorcism, in funerary rites associated with the ascent of the soul, or in other varieties of Taoist religious practice.

The "hill censer"

Some Han mirrors have astronomical or astrological patterns. The most elaborate, particularly popular during the Hsin dynasty (AD 9–25), bears the so-called TLV pattern. These angular shapes, ranged around the main band of decoration between a central square zone and the outer border band, are believed to be linked to a cosmological "chess" game called *liu-po*; the decoration also may include creatures symbolic of the four directions, immortals, and other mythical beings popular in Taoist folklore. Often the mirrors carry inscriptions, varying from a simple expression of good luck to a long dedication giving the name of the maker and referring to the Shang-fang or Imperial workshop. In the Eastern Han the Taoist elements dominated mirror design, which often includes the legendary Queen Mother of the West, Hsi Wang Mu, and her royal eastern counterpart, Tung Wang Kung. The coming of Buddhism at the end of the Han dynasty caused a decline in the use of cosmological mirrors. Mirror making, however, was revived in the T'ang dynasty.

Han glazed wares

Ceramics. Han glazed wares are chiefly of two types. Northern China saw the invention, presumably for funerary purposes only, of a low-fired lead glaze, tinted bottle-green with copper oxide, that degenerates through burial to an attractive silvery iridescence. High-fired stoneware with a thin brownish to olive glaze was still being made in Honan, but the main centre of production was already shifting to the Chekiang region, formerly known as Yüeh. Yüeh ware kilns of the Eastern Han, located at Te-ch'ing in northern Chekiang, produced a hard stoneware, often imitating the shapes of bronze vessels and decorated with impressed, bronzelike designs under a thin olive glaze. Other important provincial centres for pottery production in the Han dynasty were Ch'ang-sha (in Hunan province) and Ch'eng-tu and Chungking (in Szechwan province).

Three Kingdoms (220–280) and Six Dynasties (220–589). For 60 years after the fall of Han, China was divided among three native dynasties: the Wei in the north, Wu in the southeast, and Shu-Han in the west. It was briefly reunited under the Hsi (Western) Chin; but, in 311, Lo-yang and, in 316, Ch'ang-an fell to the invading Hsiung-nu, and before long the whole of northern China was occupied by barbarian tribes who set up one petty kingdom after another until, in 439, a Turkish tribe, the Toba, brought the region under their rule as the Pei (Northern) Wei dynasty. They established at P'eng-ch'eng (modern Ta-t'ung) in Shansi a capital that they populated by the forced immigration of tens of thousands of Chinese. The Chinese they recruited into their service civilized the Toba until they became completely Sinicized. In 495 the Wei moved their capital to Lo-yang in the heartland of ancient Chinese civilization, where they lost what little Turkish identity they still possessed. They were succeeded in 535 by other petty barbarian dynasties who held the north until the reunification of China in 581.

The barbarians adopted Buddhism as a matter of state policy, for Buddhism was an international religion with a concept of kingship that helped them to equate their earthly with their spiritual authority and thus to legitimize their control over the Chinese. Moreover, in the devastated land that was northern China in the 4th and 5th centuries, when the Confucian system was in ruins and Taoism a refuge for the few, the Buddhist doctrine of salvation through faith and good works acted as a powerful consoling and uniting force.

Buddhist missionaries and art came to Nanking by way of Indochina, but this cultural traffic did not become important before the 4th century. Although the rulers with few exceptions were weak, corrupt, or cruel and the court a maze of intrigue, it was chiefly in Nanking that the great poets, calligraphers, painters, and critics flourished, and they in turn greatly influenced the arts of the occupied north.

Architecture. After the fall of Lo-yang and Ch'ang-an, there was no more great city and palace building until

the Northern Wei moved their capital to Lo-yang in 495. There they constructed a city of great magnificence, sacked at their fall in 535. The main monuments of the 4th and 5th centuries were temples and monasteries. By the mid-6th century there were some 500 religious establishments in and around Lo-yang alone, about 30,000 in the whole of the northern realm.

Pagodas

Each Buddhist temple had its pagoda erected as a reliquary or memorial, and other pagodas dotted the city and the surrounding landscape. They have mostly disappeared, but one can get some idea of their form from reliefs at Yün-kang and from the earliest surviving pagodas at Nara in Japan. Based on an enlargement and refinement of the Han timber tower, or *lou*, they had up to 12 stories, with a projecting mast at the top ringed with metal disks. This mast was the only feature preserved from the Indian Buddhist burial or reliquary mound, the stupa, a hemispherical form that the Chinese rarely seem to have copied. The brick and stone pagodas, which were originally more Indian in form and were gradually Sinicized, are tiered structures with the stories marked by projecting string courses (horizontal bands) and architectural features borrowed from timberwork indicated in relief. The oldest surviving example is the Sung-yüeh Temple, a 12-sided stone pagoda on Mount Sung (*c.* 520–525) that is Indian in its shape and detail.

Sculpture. The earliest known Chinese Buddhist sculpture was made in the Eastern Han (at Ma-hao and P'eng-shan, in Szechwan, 2nd century AD). The first Buddhist images in gilt bronze were copies of icons brought to China from Central Asia, Afghanistan, and possibly India itself. Some are very clumsily modeled, which is not surprising when one considers that the Chinese craftsmen were copying an alien style and iconography. The finest, however, have a simple, compact charm, as can be seen in a seated Buddha (Asian Art Museum of San Francisco, The Avery Brundage Collection) copied from an Indian Gandhāra model. Bearing an inscription of 338, it is the earliest dated Chinese Buddha yet discovered.

Cave shrines

In the first centuries of Buddhist history in China, rock-cut cave temples and monastic cells rivaled timber-built courtyard temples in importance. In 460 a Chinese priest recommended to the Wei ruler that he should carve from a cliff face near the northern capital, Ta-t'ung, five huge images, dedicated to four of his Imperial predecessors and himself. These are caves XVI to XX at Yün-kang, cut in the 460s and 470s. The scale is heroic: the standing Buddha of cave XVIII and the seated Buddha of cave XX are each 14 metres high. The style is based on Central Asian models, squat in proportion and broad-shouldered, with large, square faces and staring eyes and drapery indicated in stylized repeated folds.



"Empress as Donor with Attendants," limestone relief with traces of colour, from Pin-yang Cave, Lung-men, Honan province, *c.* AD 522, Northern Wei dynasty. In The Nelson-Atkins Museum of Art, Kansas City, Mo., U.S. 1.93 m × 2.77 m.

By courtesy of The Nelson-Atkins Museum of Art, Kansas City, Missouri (Nelson Fund)

The early work at Yün-kang represents the culmination of the first phase of Buddhist sculpture in China. "Paired" caves at Yün-kang (caves VI–XII), some of which were commissioned by subsequent rulers in honour of their parents, show a progressive change in style. This style developed into an entirely different second phase, which was flat and linear and which spread to the north from Nanking. The shift of the capital southward to Lo-yang in 494 marked the beginning of massive borrowing from southern culture, as, for example, in work begun at nearby Lung-men, where the Wei emperor commissioned the Pin-yang Cave, carved between 508 and 523. Here the second phase is fully mature. The figures are flattened, the angular body almost disappearing beneath a cascade of drapery and trailing scarves; the head is elongated, the eyes are half closed, and a gentle smile touches the lips. This ethereal style, which has been compared to that characteristic of Western Romanesque sculpture, shows the influence of the courtly figure painting in 5th-century Nanking, most clearly in panels in the Pin-yang Cave depicting the emperor and empress advancing in procession with their attendants. It has been suggested that the panels may have been inspired by southern paintings brought north, such as the transportable lacquer screens in southern court style dated 484 and found in the tomb of Ssu-ma Chin-lung at Ta-t'ung.

The second phase style is eloquently displayed in stone stelae (stone slabs or pillars that were used for votive or commemorative purposes) and gilt bronze altar groups of the 5th and 6th centuries. The stelae are carved in either of two styles: slablike monoliths with figures cut in the surface or leaf-shaped mandorlas with a group, generally the Buddha and bodhisattvas, standing out in high relief. Among the most beautiful of the gilt bronzes is an image of Śākyamuni (the historical Buddha) and Prabhūtaratna of 518 (in the Guimet Museum, Paris), which shows the attenuated linear style of the second phase carried almost to the point of mannerism.

Before 550 Buddhist sculpture began to undergo a further radical change. This new third phase style, by contrast with the second phase, is solid and dignified; the standing figures are rigidly frontal and columnar and the heads massive and erect. In the bodhisattva figures, thick strands of jewelry play over the surface, making a striking contrast to the severe folds of the clinging robes; some figures are more plastic and sculptural, reflecting the Indian influence

Asian Art Museum of San Francisco, The Avery Brundage Collection



Gilt bronze seated Buddha, AD 338, Eastern Chin dynasty. In the Asian Art Museum of San Francisco, Avery Brundage Collection. Height 39.4 cm.

that was now once more penetrating into China, this time from the south, through new channels of trade and the many diplomatic and religious missions sent to Nanking by Linyi (Champa, now part of Vietnam) and Funan (approximately present-day Cambodia). From Nanking the new Indianized style spread to northern China, where it may be seen in the sculpture of the Northern Ch'i dynasty (550–577), and up the Yangtze River to Szechwan, where, at the Wan-fo Temple in Ch'eng-tu, there has been found a hoard of 6th- and 7th-century images, some of which are remarkably close to the Indian Gupta style. This third phase confrontation of Chinese and Indian art produced many styles, but at its best it attains a remarkable blend of surface richness and monumental grandeur. By the end of this phase, columnar rigidity began to give way to a more natural bearing of bodily weight, a swaying posture, and greater spatial extension (as seen in the Kuan-yin of about 570 in the Museum of Fine Arts, Boston), which became more fully developed in the Sui and T'ang dynasties.

Spread of the Indianized style

Painting. The breakdown of the Confucian system after the Han was reflected in painting and painting theory by an emphasis on Taoist and Buddhist themes and reasons for painting. This period saw the first activity by the courtier class, who painted as amateurs and who were far better remembered in the written record of the art than were their professional, artisan-class counterparts. Among the first named painting masters, Ts'ao Pu-hsing and Tai K'uei painted chiefly Buddhist and Taoist subjects. Tai K'uei was noted as a poet, painter, and musician and was one of the first to establish the tradition of scholarly amateur painting (*wen-jen hua*). He was also the leading sculptor of his day, almost the only instance in Chinese history of a gentleman who engaged in this craft.

The greatest painter at the southern court in this period was Ku K'ai-chih, an amateur painter from a family of distinguished Tung (Eastern) Wei dynasty scholar-officials in Nanking and an eccentric member of a Taoist sect. One of the most famous of his works (which survives in a T'ang dynasty copy in the British Museum) illustrates a 3rd-century didactic text "Nü-shih chen" ("Admonitions of the Court Instructress"), by Chang Hua. In this hand scroll, narrative illustration is bound strictly to the text (as if used as a mnemonic device): the advice to Imperial concubines to bear sons to the emperor, for instance, is accompanied by a delightful family group. The figures are slender and fairylike, and the line is fine and flows rhythmically. The roots of this elegant southern style, which then epitomized the highest Nanking court standard, can be traced back to Ch'ang-sha in the late Chou–early Han period, and it was later adopted as court style by the Northern Wei rulers (*e.g.*, at Lung-men) when they moved south to Lo-yang in 495. Ku K'ai-chih also was noted as a portraitist, and, among Buddhist subjects, his rendering of the sage Vimalakirti became a model for later painters.

Ku K'ai-chih

The south saw few major painters in the 5th century, but the settled reign of Wu-ti in the 6th produced a number of notable figures, among them Chang Seng-yu, who seems to have combined realism with a new freedom in the use of the brush, employing dots and dashing strokes very different from the fine precision of Ku K'ai-chih.

While all the temples of the period have been destroyed, a quantity of wall painting survives at Tun-huang in northwestern Kansu in the Caves of the Thousand Buddhas, Ch'ien-fo Tung, where there are nearly 500 cave shrines and niches dating from the 5th century onward. Early Tun-huang paintings depict chiefly incidents in the life of the Buddha, the Jātaka (stories of his previous incarnations), and such simple themes as the perils from which Avalokiteśvara (Chinese: Kuan-yin) saves the faithful. In style they show a blend of Central Asian and Chinese techniques that reflects the mixed population of northern China at this time.

Painters practicing foreign techniques were active at the northern courts in the 6th century. Ts'ao Chung-ta painted, according to an early text, "after the manner of foreign countries" and was noted for closely clinging drapery that made his figures look as though they had been drenched in water. At the end of the 6th century, a painter from Khotan (Ho-t'ien), Wei-ch'ih Po-chih-na, was active at the Sui court; a descendant of his, Wei-ch'ih I-seng, painted frescoes in the temples of Ch'ang-an using a thick impasto (a thick application of pigment) and a brush line that was "tight and strong like bending iron or coiling wire." Those foreign techniques caused much comment among the Chinese but seem to have been confined to Buddhist painting and eventually were abandoned.

The beginning of aesthetic theory in China was another product of the spirit of inquiry and introspection that was abroad in these restless years. About AD 300 a long passionate poem, "Wen Fu" ("Rhymeprose on Literature") was composed by Lu Chi on the subject of artistic creation. Also from this period, the *Wen-hsin tiao-lung* ("Literary Mind and Carving of Dragons") by Liu Hsieh (*c.* 465–*c.* 522) has long remained China's premier treatise on aesthetics. It offers insightful consideration of a wide range of chosen topics, beginning with a discussion of *wen*, or nature's underlying pattern. Set forth as central to the mastery of artistic expression are the control of "wind" (*feng*, emotional vitality) and "bone" (*ku*, structural organization).

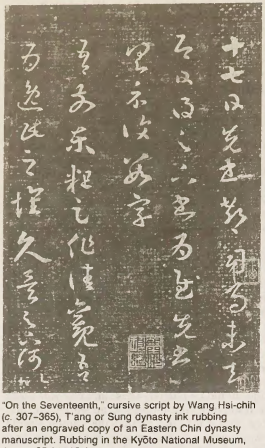
Beginning of aesthetic theory in China

In the Southern Liang dynasty critical works were written on literature and calligraphy; and, about the mid-6th century, the painter Hsieh Ho compiled the earliest work on art theory that has survived in China, the *Ku-hua p'in-lu* ("Classified Record of Painters of Former Times"). In this work he grades 27 painters in three classes, prefacing his list with a short statement of six aesthetic principles by which painting should be judged. These are: *ch'i yün sheng tung* ("spirit resonance, life-motion"), an enigmatic and much debated phrase that means that the painter should endow his work with life and movement through harmony with the spirit of nature; *ku fa yung pi* ("structural method in use of the brush"), referring to the structural power and tension of the brushstroke, alike in painting and calligraphy, through which the vital spirit is expressed; *ying wu hsiang hsing* ("fidelity to the object in portraying forms"); *sui lei fu ts'ai* (conforming to kind in applying colours); *ching ying wei chih* (planning and design in placing and positioning); and *ch'uan i mu hsieh* (transmission of ancient models by copying). Of the "six principles," the first two are fundamental, for, unless the conventional forms are brought to life by the vitality of the brushwork, the painting has no real merit, however carefully it is executed; the latter principles imply that truth to nature and tradition also must be obtained for the first two to be achieved. The six principles of Hsieh Ho have become the cornerstone of Chinese aesthetic theory down through the centuries.

The integration of spirituality and naturalism is similarly found in the short, profoundly Taoist text of the early 5th century, *Hua shan-shui hsü* ("Preface on Landscape Painting," China's first essay on the topic), attributed to Tsung Ping. Tsung suggests that if well-painted—that is, if both visually accurate and aesthetically compelling—a landscape painting can truly substitute for real nature, for, even though miniaturized, it can attract vital energy (*ch'i*) from the spirit-filled void (Tao), just as its real, material counterpart does. This interplay between macrocosm and microcosm became a constant foundation of Chinese spiritual thought and aesthetics.

Calligraphy. The fine art of writing, calligraphy, has often been held to occupy the highest place among the visual arts in China. The direct ancestor of modern writing, the script used on oracle shells and bones (*chia-ku-wen*) of the middle and late Shang dynasty, had already developed into a complex, semi-pictorial system. It gradually evolved into the large seal script (*ta-chuan shu*) seen in cast bronze inscriptions throughout the late Shang and the Chou dynasty. In the Ch'in dynasty unity was imposed by the government in the form of small seal script (*hsiao-chuan*). Perhaps because these early styles were often used for engraving inscriptions, they are all characterized by unmodulated brush lines and do not show to any advantage the expressive qualities permitted by the flexible Chinese brush. This latter quality first appeared in the Han dynasty in the form of clerical script (*li-shu*), perhaps aided by improvements in the writing brush itself. Clerical script is characterized by powerfully expanding brushstrokes, with

Script styles



"On the Seventeenth," cursive script by Wang Hsi-chih (*c.* 307–365), T'ang or Sung dynasty ink rubbing after an engraved copy of an Eastern Chin dynasty manuscript. Rubbing in the Kyōto National Museum, Japan. 25.1 × 18.4 cm.
Kyoto National Museum, photograph, Shimizu Kogeisha Co., Ltd

angular starts, turns, and stops; it can be boldly expressive whether written with the brush or carved in formal inscriptions on stone stelae. A convenient cursive version of clerical script, known as draft script (*ts'ao-shu*), also was developed, with a reduced number of strokes and considerable linkage between them. Gradually, by the end of the Han, clerical script, influenced by draft script, developed a more fluent structure and form of brushwork that has survived to this day as China's standard script (*k'ai-shu*, or *cheng-shu*). Developing parallel to the draft script was a semicursive, or "running," script (*hsing-shu*). By the early Six Dynasties period, the highly flexible standard and cursive scripts had come to be perceived as a profound artistic medium, capable of communicating fundamental personal qualities, and writing in draft script became something of a cult activity among the literati.

Among the early famous masters of calligraphy in the late Han to early Six Dynasties period were Chang Chih, Ts'ai Yung, Chung Yu, and Lu Chi. Wang Hsi-chih (*c.* 303–*c.* 361), a master of the *k'ai, hsing,* and *ts'ao* styles, especially favoured Chung Yu's fluid style but studied the past broadly and achieved the first great historical synthesis of styles. His sons Wang Hsien-chih and Wang Hui-chih inherited his talents (the whole family seems also to have been engaged in occult Taoism, in which their calligraphy was used for mystic trance-writing). For a while, Wang Hsien-chih's more suave manner of calligraphy surpassed his father's more spontaneous art in popularity. But the father was idolized by the second T'ang dynasty emperor, T'ai-tsung, who had his works copied at the 7th-century court for widespread distribution and took almost all the originals to the tomb with him, including Wang's *Lan-ting hsü* ("Orchid Pavilion Preface"), the most hallowed work of visual art in all East Asian history; and for over a millennium Wang Hsi-chih has been imitated by nearly every Asian schoolchild with a brush. Wang Hsi-chih and his followers founded a southern style of calligraphy that is elegant and graceful and that contrasted with the somewhat archaic style in the north at that time. The latter, known especially from stelae inscriptions, perpetuated the bold and angular power of Han clerical script and belatedly came to rival Wang Hsi-chih's influence in the hands of 8th-century T'ang masters such as Chang Hsü, who effected a second great historical synthesis.

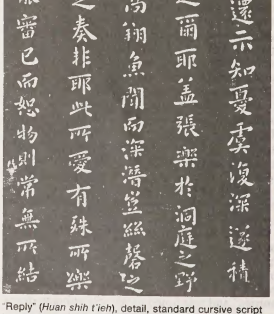
Ceramics. The increase in population in the lower Yangtze valley was a great stimulus to the pottery industry in the Six Dynasties. Kilns in Chekiang (the old kingdom of Yüeh) were producing a stoneware with an olive brown or greenish glaze. Examples of Yüeh ware jars, ewers, pitchers, and other grave goods have been found in 3rd- and 4th-century tombs in the Nanking region. They were made chiefly at Shao-hsing, at Shang-lin Lake, and at Te-ch'ing, north of Hang-chou, which also produced a stoneware with a glossy black glaze. During the Six Dynasties potters freed themselves from the influence of bronze design and produced shapes more characteristic of pottery. Yüeh stoneware

While most of the Chekiang wares are plain or simply decorated, "northern celadon," produced in Hopeh and Honan in the 6th century, is exotic in style, reflecting the taste of Turkish rulers and other cultural contacts with western Asia. Heavy funerary jars are adorned with acanthus and lotus leaves, and flowers and round decorative plaques are molded or applied to the surface in imitation of Sāsānian repoussé metalwork. Tomb figurines of this period are often made of dark gray earthenware and unglazed, though sometimes they are painted.

Sui (581–618) and T'ang (618–907) dynasties. The founding of the Sui dynasty reunited China after more than 300 years of fragmentation. The second Sui emperor engaged in unsuccessful wars and vast public works, such as the Grand Canal linking the north and south, that exhausted the people and caused them to revolt. The succeeding T'ang dynasty built a more enduring state on the foundations the Sui rulers had laid, and the first 130 years of the T'ang was one of the most prosperous and brilliant periods in the history of Chinese civilization. The empire now extended so far across Central Asia that for a while Bukhara and Samarkand were under Chinese control, the Central Asian kingdoms paid China tribute, and Chinese cultural influence reached Korea and Japan. Ch'ang-an became the greatest city in the world: its streets were filled with foreigners, and foreign religions—including Zoroastrianism, Buddhism, Manichaeism, Nestorianism, Christianity, Judaism, and Islām—flourished. This confident cosmopolitanism is reflected in all the arts of this period.

The splendour of the dynasty reached its peak between 712 and 756 under Hsüan-tsung (or Ming-huang), but before the end of his reign a disastrous defeat lost Central

© The Field Museum, Chicago (neg. no. A100889)



"Reply" (*Huan shih t'ieh*), detail, standard cursive script by Chung Yu, dated 221, Wei dynasty (ink rubbing from a 10th-century engraved copy). Rubbing in The Field Museum, Chicago. Height 27.9 cm.

Decline of the T'ang

Asia to the advancing Arabs, and the rebellion of General An Lu-shan in 755 almost brought down the dynasty. Although the T'ang survived another 150 years, its great days were over; and, as the empire shrank and the economic crisis deepened, the government and people turned against foreigners and foreign religions. In 845 all foreign religions were briefly but disastrously proscribed; temples and monasteries were destroyed or turned to secular use, and Buddhist bronze images were melted down. Today the finest Buddhist art and architecture in the T'ang style is to be found not in China but in the 8th-century temples at Nara in Japan. While the ancient heartland of Chinese civilization in the Honan-Shensi area sank in political and economic importance, the southeast became ever more densely populated and prosperous, and in the last century of the T'ang it was once again the cultural centre of China, as it had been in the Six Dynasties.

Architecture. The Sui capital, Ta-hsing, was designed in 583 on Imperial order by the great architect Yü-wen K'ai; renamed Ch'ang-an, it was further developed by the T'ang after 618. This vast city, six times the size of present-day Sian, was laid out in nine months on a grid plan, with eastern and western markets and the Imperial City placed in the central northern section, a plan later followed at Peking. In 634 T'ang T'ai-tsung built a new palace, the Ta-ming Palace, on higher ground just outside the city to the northeast. The site of the Ta-ming Palace, which became the centre of court life during the glittering reigns of Kao-tsung (649–683) and Hsüan-tsung (712–756), has recently been partly excavated. Remains have been found of two great halls, Han-yüan Hall, with its elevated corridors extended like huge arms toward overlapping triple towers (foreshadowing the later Japanese Phoenix Hall at Uji and the Wu Gate at Peking), and the Lin-te Hall; marble flagstones and bases of 164 columns of the latter give some indication of its splendour. Lost marvels of Sui-T'ang palace architecture include Yü-wen K'ai's rotating pavilion in the Sui palace, which could hold 200 guests, and the 90-metre-high state Ming-t'ang ("Spirit Hall") built for China's only reigning empress, the usurper Wu-hou (or Wu Tse-t'ien, who changed the name of the dynasty from T'ang to Chou during her reign from 690 to 705). Surviving murals from Buddhist caves at Tun-huang and excavated royal tombs near Ch'ang-an provide a graphic record of T'ang architecture, its taste for multistory elevation, tall towers, and elaborate elevated walkways, its somewhat garish use of coloured-tile building surfaces, and its integration of architecture with gardens, ponds, and bridges.

The Sui-T'ang period saw some of China's most lavish royal tomb building, before the onset of a relative modesty in the Sung and a decline of qualitative standards in later periods. Excavated royal tombs at Ch'ang-ling, north of the capital, include three built for close relatives of Wu-hou who were degraded or executed by her on her way to the throne and reburied amid much pomp and splendour in 706 after the restoration of the T'ang royal lineage. In each, the subterranean sepulchre is surmounted by a truncated pyramidal tumulus and is approached through a sculpture-lined "spirit way" (*ling-tao*). Inside, painted corridors and incised stone sarcophagi provide a lingering record of T'ang splendour, with colourful renderings of palatial settings, foreign diplomats, servants-in-waiting, and recreation at polo and the hunt. Along the corridor, niches that had served temporarily as ventilating shafts are stuffed with ceramic figurines, riders and entertainers, T'ang horses and other fabulous animals, mostly done in bold tricolour glazes. The corridor leads to two domed vaults serving as an antechamber and burial hall. The tombs of some T'ang rulers were so grand that artificial tomb mounds no longer sufficed, and funerary caverns were carved out beneath large mountains. The huge tomb of Emperor Kao-tsung and his empress, Wu-hou (China's only joint burial of rulers), at Ch'ang-ling, has yet to be excavated but appears to be intact.

The Sui and the first half of T'ang were great periods of temple building. The first Sui emperor distributed relics throughout the country and ordered that pagodas and temples be built to house them, and the early T'ang monarchs were equally lavish in their foundations. Apart from masonry pagodas, however, very few T'ang temple buildings have survived. The oldest yet identified is the main hall of Nan-ch'an Temple at Wu-t'ai in northern Shansi (before 782); the largest is the main hall of nearby Fo-kuang Temple (857). The Great Buddha Hall (Daibutsu-den, 752) of the Tōdai Temple (Tōdai-ji) at Nara in Japan, 88 metres long and 51.5 metres wide, built in the T'ang style, is today the largest wooden building in the world; however, it is small compared with the lost T'ang temple halls of Lo-yang and Ch'ang-an.

T'ang and later pagodas show little of the Indian influence that was so marked on the Sung-yüeh Temple pagoda. T'ang wooden pagodas have all been destroyed, but graceful examples survive at Nara, notably at Hōryū Temple, Yakushi Temple, and Daigō Temple. Masonry pagodas include the seven-story, 58-metre-high Ta Yen T'a, or Great Wild Goose Pagoda, of Tz'u-en Temple in Ch'ang-an, on which the successive stories are marked by corbeled cornices, and timber features are simulated in stone by flat columns, or pilasters, struts, and capitals.

Masonry pagodas

T'ang cave-temples at Tun-huang were increasingly Sinicized, abandoning the Indianesque central pillar, the circumambulated focus of worship which in Six Dynasties caves was sculpted and painted on all four sides with Buddhist paradises; in the T'ang, major Buddhist icons and paradise murals were moved to the rear of an open chamber and given elevated seating, much like an emperor enthroned in his palace or like any Chinese host.

Sculpture. Although the Sui emperors were great patrons of Buddhism, there were no major changes in style during this brief period. T'ang T'ai-tsung was hostile to Buddhism, but his successor, Kao-tsung, and Empress Wu were lavish in their endowments. Under Kao-tsung, the principal cave shrine, Feng-hsien Temple, was carved out at Lung-men between 672 and 675. The central figure, cut almost in the round, is the cosmic Buddha Vairocana (one of the Five Celestial Buddhas), whose worship had recently been introduced with the doctrines of the Esoteric Chen-yen sect. The figure, 11 metres high, is compact, restrained, and somewhat severe in conception; the head and body are massively yet delicately modeled, and the features and drapery are sharply carved, as though painted with a fine brush.

The next 50 years saw the climax of the fourth phase of Chinese Buddhist sculpture, in which Indian mass and Chinese rhythmic line were triumphantly united and rec-

Courtesy of The Arthur M. Sackler Museum, Harvard University Art Museums, bequest of Grenville L. Winthrop



Buddha from cave XXI, T'ien-lung Shan, Shansi province; late 7th–early 8th century, T'ang dynasty. Limestone with traces of paint. In The Arthur M. Sackler Museum, Cambridge, Mass., U.S. Height 1.1 m.

onciled. The high point came about 690–710, in the early caves at T'ien-lung Shan, carved under the patronage of Empress Wu. The largest and finest bronze pieces in this style are a bronze Yakushi flanked by bodhisattvas (Yakushi Temple, Nara); no examples of comparable quality have survived in China.

The mature style of the 8th century is well displayed in later caves at T'ien-lung Shan, carved for the emperor Hsüan-tsung. Here the modeling is much fuller and more fleshy, the poses more exaggerated, and the drapery sweeps over the partly exposed body in an emphatic manner very different from the restrained nobility of the 7th century. (This "baroque" style was an expression not only of Indian sculptural influences but also of the confident materialism of high T'ang culture, when the spiritual message of Buddhism seemed less important than the lavish endowment of temples and images and the display of religious extravagance that was one of the chief causes of the anti-Buddhist campaign of 845.

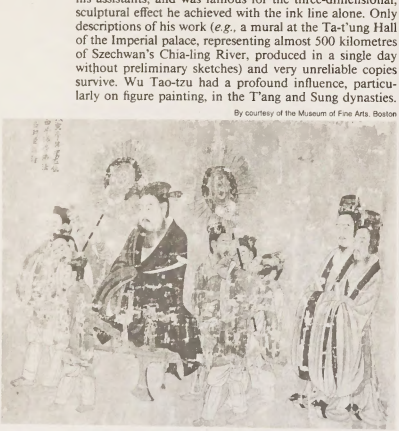
"Baroque" style of sculpture

Painting. The patronage of the Sui and T'ang courts attracted painters from all over the empire. Yen Li-pen, who rose to high office as an administrator, finally becoming a minister of state, was also a noted 7th-century figure painter. His duties included painting historical scrolls, notable events past and present, and portraits, including those of foreigners and strange creatures brought to court as tribute, to the delight of his patron, T'ai-tsung. Yen Li-pen painted in a conservative style with a delicate, scarcely modulated line. Part of a scroll depicting 13 emperors from Han to Sui (in the Museum of Fine Arts, Boston) is attributed to him.

The royal tombs near Sian (706) show the emergence of a more liberated tradition in brushwork that came to the fore in mid- to late-8th-century painting, as it did in the calligraphy of Chang Hsü, Yen Chen-ch'ing, and other master writers. The greatest brush master of T'ang painting was the 8th-century artist Wu Tao-hsüan, who was also called Wu Tao-tzu and who not only enjoyed a career at court but had sufficient creative energy to execute, according to T'ang records, some 300 wall paintings in the temples of Lo-yang and Ch'ang-an. His brushwork, in contrast to that of Yen Li-pen, was full of such sweeping power that crowds would gather to watch him as he worked. He painted chiefly in ink, leaving the colouring to his assistants, and was famous for the three-dimensional, sculptural effect he achieved with the ink line alone. Only descriptions of his work (*e.g.*, a mural at the Ta-t'ung Hall of the Imperial palace, representing almost 500 kilometres of Szechwan's Chia-ling River, produced in a single day without preliminary sketches) and very unreliable copies survive. Wu Tao-tzu had a profound influence, particularly on figure painting, in the T'ang and Sung dynasties.

Influence of Wu Tao-tzu

By courtesy of the Museum of Fine Arts, Boston



Portrait of Ch'en Hsüan Ti carried by his servants, detail from "Portraits of the Emperors," hand scroll attributed to Yen Li-pen (7th century), T'ang dynasty. Colour on silk. In the Museum of Fine Arts, Boston. 10.96 m × 46 cm.



Polo player, detail of a mural from the tomb of Li Hsien, the crown prince Chang-huai, at Ch'ien-hsien near Sian, Shensi province, 706, T'ang dynasty. In the Shensi Provincial Museum, Sian. Detail 190 × 175 cm.

Wang Lu/ChinaStock Photo Library

Figure painters who depicted court life in a careful manner derived from Yen Li-pen rather than from Wu Tao-tzu included Chang Hsüan and Chou Fang. Eighth-century royal tomb murals and Tun-huang Buddhist paintings demonstrate the early appearance and widespread appeal of styles that these court artists helped later to canonize, with individual figures (especially women) of monumental, sculpturesque proportion arranged upon a blank background with classic simplicity and balance.

Horses played an important role in T'ang military expansion and in the life of the court; riding was a popular recreation, and even the court ladies played polo. Horses also had become a popular subject for painting, and one of the emperor Hsüan-tsung's favourite court artists was the horse painter Han Kan. The other great horse painting master was the army general Ts'ao Pa, said by the poet Tu Fu to have captured better the inner character of his subjects and not just the flesh. Most later horse painters claimed to follow Han Kan or Ts'ao Pa, but the actual stylistic contrast between them was already reported in Northern Sung times as no longer distinguishable and today is hardly understood.

The more than three centuries of the Sui and T'ang were a period of progress and change in landscape painting. The early 7th- and 8th-century masters Chan Tzu-ch'ien, Li Ssu-hsün, and the latter's son Li Chao-tao developed a style of landscape painting known as *ch'ing-lü-pai* ("green, blue, white"), or *chin-pi shan-shui* ("gold-blue-green"), in which mineral colours were applied to a composition carefully executed in fine line to produce a richly coloured effect. Probably related to Central Asian painting styles of the Six Dynasties period and associated with the jeweled-paradise landscapes of the Taoist immortals, this "blue-and-green" type readily appealed to the T'ang court's taste for international exotica, religious fantasy, and boldly decorative art. A painting in this technique, known as "Ming-huang's Journey to Shu" (that is, Szechwan; in the National Palace Museum, Taipei), reflects what is considered to be the style of Li Chao-tao, although it is probably a later copy. This style gradually crystallized as a courtly and professional tradition, in contrast to the more informal calligraphic ink painting of the literati.

"Blue-and-green" style of landscape painting

The traditional founder of the school of scholarly landscape painting (*wen-jen hua*) is Wang Wei, an 8th-century scholar and poet who divided his time between the court at Ch'ang-an, where he held official posts, and his country estate of Wang Ch'uan, of which he painted a panoramic composition preserved in later copies and engraved on stone. Among his Buddhist paintings, the most famous

was a rendering of the Indian sage Vimalakirti, who became, as it were, the "patron saint" of Chinese Buddhist intellectuals. Wang Wei sometimes painted landscapes in colour, but his later reputation was based on the belief that he was the first to paint landscape in monochrome ink. He was said to have obtained a subtle atmosphere by "breaking the ink" (*p'o-mo*) into varied tones. The belief in his founding role, fostered by later critics, became the cornerstone of the philosophy of the *wen-jen hua*, which held that a man could not be a great painter unless he was also a scholar and a gentleman.

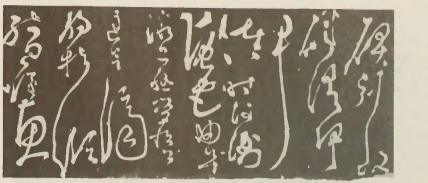
More adventurous in technique was the somewhat eccentric late-8th-century painter Chang Tsao, who produced dramatic tonal and textural contrasts, as when he painted simultaneously, with one brush in each hand, two branches of a tree, one moist and flourishing, the other desiccated and dead. This new freedom with the brush was carried to extremes by such painters of the middle to late T'ang as Wang Hsia (or Wang Mo) and Ku K'uang, southern Chinese Taoists who "splashed ink" (also transliterated as *p'o-mo* but written with different characters than "broken ink") onto the silk in a manner suggestive of 20th-century Action painters. The intention of these ink-splashers was philosophical and religious as well as artistic: it was written at the time that their spontaneous process was designed to imitate the divine process of creation. Their semi-finished products, in which the artistic process was fully revealed and the subject matter had to be discerned by the viewer, suggested a Taoist philosophical skepticism. These techniques marked the emergence of a trend toward eccentricity in brushwork that had free rein in periods of political and social chaos. They were subsequently employed by painters of the southern "sudden" school of Ch'an (Zen) Buddhism, which held that enlightenment was a spontaneous, irrational experience that could be suggested in painting only by a comparable spontaneity in the brushwork. Ch'an painting flourished particularly in Ch'eng-tu, the capital of the petty state of Shu, to which many artists went as refugees from the chaotic north in the last years before the T'ang dynasty fell, among them the eccentric artists Kuan-hsiu and Shih K'o.

Trend toward eccentricity in brushwork

Calligraphy. As T'ang painting moved in the direction of greater linear expressiveness, it came increasingly under the sway of developments in calligraphy, not only in technique and style but in aesthetic theory as well. The Sinicization of the Turkic rulers of the Northern Wei by the late 5th century and the reunification of China under the Sui and T'ang a century later paved the way for an infusion of southern taste at the court in the north and, gradually, for a synthesis of southern and northern styles. The passionate accumulation and distribution through careful copywork of Wang Hsi-chih's surviving manuscripts by the second T'ang dynasty emperor, T'ai-tsung, represents this infusion and was accompanied by widespread calligraphic activity at the court, where Ou-yang Hsün, Yü Shih-nan (a pupil of Wang Hsi-chih's 7th-generation descendant, the priest Chih-yung), and Ch'u Sui-liang were among the many brush masters of lasting fame who were gathered to serve as copyists for the emperor and the nation. T'ai-tsung himself was a notable writer, as was his descendant of the next century, Hsüan-tsung.

It was during Hsüan-tsung's reign that Chinese calligraphy reached a third definitive period, comparable to the Ch'in unification of seal script and the Wang family's synthesis of earlier styles. The three greatest masters involved in synthesizing southern and northern tendencies, blending fluid brush movement and expressive power, were Chang Hsü (known for his "delirious cursive" writing) and his pupils Yen Chen-ch'ing and the monk Huai-su (the latter practicing a "wild cursive"—*k'uang-ts'ao*—style). They helped establish a new aesthetic standard that would soon pervade painting as well as calligraphy, more fully than ever before realizing in visual form the ancient aesthetic principles of "natural" emotionality and "sincerity" in rendering. Their preference for rugged, intentionally awkward, or altogether surprising forms was quite different from the elegant polish of the earlier, southern Chinese styles; and their emphasis on expressing character and emotion was handed down through Liu Kung-ch'üan of the next century and enshrined as the dominant manner of the Sung dynasty by Su Shih, Huang T'ing-chien, and Mi Fu of the 11th century.

T'ang masters of calligraphy



"Thousand-Character Essay," detail, cursive script by Chang Hsü (*c.* 700–750), T'ang dynasty (ink rubbing from a fragmentary stone copy). Rubbing in The Field Museum, Chicago. Detail 27.9 × 74.9 cm.
© The Field Museum, Chicago (neg. no. A100879)

Ceramics. After the comparative sterility of the Six Dynasties, this was a great period in the development of Chinese pottery. Although a white porcelain perfected early in the 7th century is called Hsing *yao* (Hsing "ware") because of a reference to white porcelain of Hsing-chou in the 9th-century essay *Ch'a Ching* ("Tea Classic") by Lu Yü, as yet no kilns have been found there. Kilns near Ting-chou in Hopeh, however, were at this time already producing a fine white porcelain, ancestor of the famous Ting ware of the Northern Sung. Late-7th- and 8th-century ceramists in northern China, working primarily at kilns at T'ung-ch'uan near Ch'ang-an and at Kung-hsien in Honan province, also developed "three-colour" (*san ts'ai*) pottery wares and figurines that were slipped and covered with a low-fired lead glaze tinted with copper or ferrous oxide in green, yellow, brown, and sometimes blue; the bright colours were allowed to mix or run naturally over the robust contour of these vessels, which are among the finest in the history of Chinese pottery. Northern Chinese kilns in Shensi also produced a stoneware with a rich black glaze, and a type of celadon was made north of Sian, in Shensi. The northern Chinese potters borrowed shapes and motifs from western Asia even more freely than had their 6th-century predecessors; foreign shapes include the amphora, bird-headed ewer, and rhyton; foreign motifs include hunting reliefs, floral medallions, boys with garlands or swags of vines, and Buddhist symbols adapted and applied with characteristic T'ang confidence. Some forms were borrowed from metalwork or glassware.

Tomb figurines were produced in such enormous quantities that attempts were made through sumptuary laws to limit their number and size, but they met with little success. The figurines were made, generally in molds, of earthenware covered with slip and painted or glazed or both. Among the human figures are servants and actors, female dancers, and musicians of exquisite grace. The 7th-century figurines are slender and high-waisted, those of the 8th century are increasingly rotund and round-faced, reflecting a change in fashion. There are also many figurines of Central Asian grooms and Semitic merchants, whose deep-set eyes and jutting noses are caricatured. Of the camels and horses, the most remarkable are glazed camels bearing on their backs a group of four or five singers and musicians. After the middle of the 8th century, there was a sharp decline both in the quantity and in the quality of northern China tomb wares and figurines.

The great southward movement of population in the T'ang dynasty stimulated the development of many new kilns. Celadons were now made in Chiung-lai (Szechwan), Ch'ang-sha (Hunan), and several areas of Kwangtung and Fukien. A kiln producing whitewares was active at Chi-chou in Kiangsi, and at Ching-te-chen in the same province two kilns were producing celadons and whitewares. From these humble beginnings, Ching-te-chen was destined to become, in the Ming and Ch'ing dynasties, the largest pottery factory in the world. In the *Ch'a Ching*, the celadons of Yüeh-chou in Chekiang are ranked for their jadelike quality first among the wares suitable

Yüeh celadons

for tea drinking, followed by the silvery Hsing ware.

Metalwork. The metalwork and jewelry of the T'ang period are witness to the wealth and cosmopolitanism of T'ang culture. Silver and silver-gilt ritual objects of many kinds have been found in T'ang tombs, including vases, ewers, dishes, bowls, wine cups, and incense burners. While some shapes are traditional Chinese, others are foreign. Alms bowls, *kuṇḍikā* water bottles (Buddhist ewers used in the initiation of monks), and reliquaries arrived with Buddhism from India; a polylobed cup, platter with relief decoration, ewer, and handled cup are of Sāsānian origin; the rhyton drinking cup is ultimately Greek. Among decorative motifs, the pearl band, cloud volute, curl border, dragons, and phoenixes are Chinese; hunting scenes, vine scrolls, rosettes, and opposed birds or animals in roundels are Sāsānian; and from India came Buddhist themes, lotus scrolls, and much of the floral ornament.

Sui and T'ang bronze mirrors

Many of the finest Sui and T'ang bronze mirrors were made at Yang-chou on the Grand Canal. Decoration, cast in high relief or repoussé, includes auspicious motifs, Taoist scenes, and the 12 zodiacal animals. The "lion and grape" mirror was very popular in the 7th and 8th centuries, for it embodied both Chinese directional and "Five Phases" symbolism, and lions and grapes were also potent symbols in the Manichaean faith introduced by the Uighur from Central Asia. This type of mirror disappeared abruptly with the persecution of the Manichaeans in 843.

Decorative arts: collection in the Shōsō-in treasure house. The Shōsō-in treasure house, a timber structure in Nara, Japan, was built to receive the personal treasures bequeathed to the Tōdai Temple by the emperor Shōmu, who died in 756. While subsequent deposits gradually added to the collection, the original gift embraced more than 600 items, which included Buddhist ritual objects, furniture, musical instruments, textiles, metalwork, lacquerwork, cloisonné, glassware, pottery, painted screens, calligraphy, and maps. Many of these pieces must have been made in Japan, but they are for the most part typically T'ang in style and decoration. This collection of T'ang-style decorative arts and crafts is the greatest in the world. Its importance lies in the fact that it is exactly datable to 756 or earlier, nearly all the pieces are in excellent condition, and they include types of decoration and technique of which no examples have survived in China.

Five Dynasties (907–960) and Ten Kingdoms (902–978). At the fall of the T'ang, northern China, ruled by five short-lived dynasties, plunged into a state of political and social chaos. The corrupt northern courts offered little support to the arts, although Buddhism continued to flourish until persecution in 955 destroyed much of what had been created in the 110 years since the previous anti-Buddhist campaign. The 10 independent kingdoms that ruled various parts of southern China, though no more enduring, offered more enlightened patronage. At first, the Ch'ien (Former) Shu (with its capital at Ch'eng-tu), then, for a longer period, the kingdoms of the Nan (Southern) T'ang (with the capital at Nanking) and Wu-Yüeh (with its capital at Hang-chou) were centres of comparative peace and prosperity. Li Hou-chu (or Li Yü), the last ruler of the Southern T'ang, was a poet and liberal patron at whose court the arts flourished more brilliantly than at any time since the mid-8th century. Not only were the southern courts at Ch'eng-tu and Nanking leading patrons of the arts but they also began formalizing court sponsorship of painting by organizing a centralized atelier with an academic component and by granting painters an elevated bureaucratic stature—policies that would be followed or modified by subsequent dynasties.

The greatest of the northern painters

Painting. In northern China only a handful of painters were working. The greatest of them, Ching Hao, who was active from about 910 to 950, spent much of his life as a recluse in the T'ai-hang Mountains of Shansi. No authentic work of his survives, but he seems from texts and later copies to have created a new style of landscape painting. Boldly conceived and executed chiefly in ink with firmness and concentration, his precipitous crags, cleft with gullies and rushing streams, rise up in rank upon rank to the top of the picture. For 150 years before his time the centre of landscape painting had been in the southeast, and Ching Hao's importance lies in the fact that he both revived the northern spirit and created a type of painting that became the model for his follower Kuan T'ung and the classic northern masters of the early Sung period, Li Ch'eng and Fan K'uan. An essay on landscape painting, "Pi-fa chi" ("Notes on Brushwork"), attributed to Ching Hao, sets out the philosophy of this school of landscape painting, one that was consistent with newly emergent Neo-Confucian ideals. Painting was to be judged both by its visual truthfulness to nature and by its expressive impact. Consistent with this, in all the major schools of Sung landscape painting that followed, artists would render with remarkable accuracy their own regional geography, letting it serve as a basis for their styles, their emotional moods, and their personal visions.

In contrast to the stark drama of this northern style, landscapes associated with the name of Tung Yüan, who held a sinecure post at the court of Li Hou-chu in Nanking, are broad, almost Impressionist in treatment. The coarse brushstrokes (known as "hemp-fibre" texture strokes), dotted accents ("moss dots"), and wet ink washes of his monochrome style, said to be derived from Wang Wei, suggest the rounded, tree-clad hills and moist atmosphere of the Chiang-nan ("South of the River") region. The contrast between the firm brushwork and dramatic compositions of such northern painters as Ching Hao and his followers and the more relaxed and spontaneous manner of Tung Yüan and his follower Chü-jan laid the foundation for two distinct traditions in Chinese landscape painting that have continued up to modern times. The style developed by Tung Yüan and Chü-jan became dominant in the Ming and Ch'ing periods, preferred by amateur artists because of its easy reduction to a calligraphic mode, its calm and understated compositional nature, and its regional affiliation.

While the few figure painters in northern China, such as Hu Huai, characteristically recorded hunting scenes, the southerners, notably Ku Hung-chung and Chou Wen-chü, depicted the voluptuous, sensual court life under Li Hou-chu. A remarkable copy of an original work by Ku Hung-chung depicts the scandalous revelries of the minister Han Hsi-tsai. Chou Wen-chü was famous for his pictures of court ladies and musical entertainments, executed with a fine line and soft, glowing colour in the tradition of Chang Hsüan and Chou Fang.

Flower painting

Flower painting, previously associated chiefly with Buddhist art, came into its own as a separate branch of painting in the Five Dynasties. At Ch'eng-tu, the master Huang Ch'üan brought to maturity the technique of *mo-ku hua* ("boneless painting"), in which he applied light colours with delicate skill, hiding the intentionally pale underdrawing and seeming thereby to dispense with the usually dominant element of strong brush outline. His great rival, Hsü Hsi, working for Li Hou-chu in Nanking, first drew his flowers in ink in a bold, free manner suggestive of the draft script, *ts'ao-shu,* adding a little colour afterward. Both men established standards that were followed for centuries afterward. Because of its reliance on technical skill, Huang Ch'üan's naturalistic style (also referred to as *hsieh-sheng,* or "lifelike painting"), was mainly adopted by professional painters, while the scholars admired the calligraphic freedom of Hsü Hsi's style (referred to as *hsieh-i,* or "painting the idea"). After the founding of the Sung, *hsieh-sheng* artists from Szechwan, including Huang Ch'üan and his sons Huang Chü-ts'ai and Huang Chü-pao, traveled to the new court at Pien-ching (K'ai-feng), where they established a tradition that dominated the Northern Sung period. Hsü Hsi found greater favour in the Yüan, Ming, and Ch'ing periods.

Ceramics. The confusion of northern China under the Five Dynasties was not conducive to development of the pottery industry, and some types such as the T'ang three-colour wares went out of production completely. White porcelain and black glazed stonewares, however, continued into the Sung dynasty. By contrast, the flourishing southern courts and the massive increase in the population of southeastern China were a great stimulus to the craft. A large complex of kilns that had been established at Yü-yao, around the Shang-lin Lake in Chekiang, which

lay in the territory of the kingdom of Wu-Yüeh, sent its finest celadons to the court of Li Hou-chu until his realm fell to the Sung in 978; after that they were sent as tribute to the Sung court at Pien-ching. The finest pieces, with decoration carved in the clay body under a very pale olive green glaze, were called by 10th-century writers, who implied that the colour was produced in imitation of jade, *pi-se yao* ("secret," or "reserve, colour ware"). It is not known whether this referred to a secret process or to the fact that the ware was reserved for the court.

Sung (960–1279), Liao (907–1125), and Chin (1115–1234) dynasties. Although reunited and ably ruled for well over a century by the first five Sung emperors, China failed to recover the northern provinces from the barbarians. A Khitan tribe, calling their dynasty Liao, held all of northeastern China until 1125, while the Hsi (Western) Hsia held the northwest, cutting off Chinese contact with western and Central Asia. From the new capital, Pien-ching, the Sung rulers pursued a pacific policy, buying off the Khitan and showing unprecedented toleration at home. While it brought Chinese scholarship, arts, and letters to a new peak of achievement, this policy left the northern frontiers unguarded. When in 1114 the Juchen Tatars in the far northwest revolted against the Khitan, the Chinese army helped the rebels destroy their old enemy. The Juchen then turned on the Sung, invaded China, beseiged the capital in 1126, and took the emperor Ch'in-tsung, the emperor emeritus Hui-tsung, who had recently abdicated, and the Imperial court prisoner, establishing their own dynasty, the Chin, with their capital at the city later to be called Peking. The remnants of the Sung court fled to the south in 1127 and, after several years of wandering, established their "temporary" capital at the beautiful city of Hang-chou. The Nan (Southern) Sung never seriously tried to recover the north but enjoyed the beauty and prosperity of their new home, while the arts continued to flourish in an atmosphere of humanity and tolerance until the Mongols entered China in the 13th century and swept all before them. In 1234 they destroyed the Juchen Tatars, and, although the Chinese armies resisted valiantly, Hang-chou fell in 1276. Three years later a loyal Sung minister drowned himself and the young emperor.

Fall of the Northern Sung

The Pei (Northern) Sung was a period of reconstruction and consolidation. Pien-ching was a city of palaces, temples, and tall pagodas; Buddhism flourished, and monasteries and temples once again multiplied. The Sung emperors attracted around them the greatest literary and artistic talent of the empire, and something of this high culture was carried on by their successors of Liao and Chin. The atmosphere at the Southern Sung court in Hang-chou, while if possible more refined and civilized, was clouded by the loss of the north; the temptation to enjoy the delights of Hang-chou and neglect their armies on the frontier turned men in on themselves. The power and confidence are gone from Southern Sung art; instead it is imbued with an exquisite sensibility, a sometimes poignant romanticism that seems for all its beauty to contain the seeds of the disaster that befell China in the 13th century.

Sung interest in history and a revival of the Classics were matched by a new concern with the tangible remains of China's past. This was the age of the beginning of archaeology and of the first great collectors and connoisseurs. One of the most enthusiastic of these was the Northern Sung emperor Hui-tsung (1100–1125/26), whose passion for the arts blinded him to the perils that threatened his country. Hui-tsung's sophisticated antiquarianism reflects an attitude that became an increasingly important factor in Chinese art. He collected and cataloged pre-Ch'in bronzes and jades while the palace studios turned out close replicas and archaistic emulations of both media. Building his royal garden, the Ken-yüeh, was said to have nearly bankrupted the state, as gigantic garden stones hauled up by boat from the south closed down the Grand Canal for long periods. He was also the most distinguished of all Imperial painting collectors, and the catalog of his collection (the *Hsüan-ho hua-p'u*, encompassing 6,396 paintings by 231 painters) remains a valuable document for the study of early Chinese painting. (Part of the collection passed into the hands of the Chin conquerors, and the remainder was scattered at the fall of Pien-ching.) Hui-tsung also elevated to new heights the recent process of bureaucratizing court painting, with entrance examinations modeled on civil service norms, with ranks and promotions like those of scholar-officials, and with regularized instruction sometimes offered by the emperor himself as chief academician. The favours granted throughout the Sung to lower-class artisans at court incurred the ire of aristocratic courtiers and provided stimulus for the rise of the amateur painting movement among these scholar-officials (*shih-ta-fu hua*), which ultimately became the literati painting mode (*wen-jen hua*) that dominated most of Yüan, Ming, and Ch'ing history.

Bureaucratization of court painting

Architecture. The Sung capital, Pien-ching, grew to a great city, only to be burned by Juchen Tatars in 1127, just after the work was completed. Nothing survives today, but some idea of the architecture of the city is given by a remarkably realistic hand scroll, "Going Up the River at Ch'ing-ming Festival Time," painted by the 12th-century court artist Chang Tse-tuan (whether painted before or after the sacking is uncertain). From contemporary accounts, Pien-ching was a city of towers, the tallest being a pagoda 110 metres high, built in 989 by the architect Yü Hao to house a relic of the Indian emperor Aśoka. Palaces and temples were at first designed in the T'ang tradition, sturdy and relatively simple in detail though smaller in scale. The plan and grouping of the elements, however, became progressively more complex; temple halls were often built in two or three stories, and structural detail became more elaborate.

Wan-go H.C. Weng Photo Collection, New York



"Going Up the River at Ch'ing-ming Festival Time," detail of a hand scroll by Chang Tse-tuan (12th century), Sung dynasty. Ink and colour on silk. In the Palace Museum, Peking. 24.8 cm × 528 cm.

The style of the 10th century is exemplified in the Kuan-yin Hall of the Tu-le Temple at Chi-hsien, Hopeh province, built in 984 in Liao territory. A two-story structure with a mezzanine that projects to an outer balcony, the hall is effectively constructed of three tiers of supporting brackets. It houses a 16-metre-high 11-headed clay sculpture of Kuan-yin, the largest of its kind in China, placed majestically beneath a central canopy. From the 11th century, the finest surviving buildings are the main hall and library of the Hua-yen Temple in the Liao capital at Ta-t'ung (Shansi), which was accorded the right to house images of the Liao emperors, installed in 1062. The library, perhaps the most intricate and perfectly preserved example of the architecture of the period, was completed in 1038.

The new Sung style is characterized by a number of distinct features. The line of the eaves, which in T'ang architecture of northern China was still straight, now curves up at the corners, and the roof has a pronounced sag-

ging silhouette. The bracket cluster (*tou-kung*) has become more complex: not only is it continuous between the columns, often including doubled, or even false, cantilever arms (or "tail-rafters," *hsia-ang*), which slant down from the inner superstructure to the bracket, but also a great variety of bracket types may be used in the same building (56 different types are found in the five-story wooden pagoda built in 1056 at the Fo-kung Temple in Ying-hsien, Shansi province). The tail-rafter, hitherto anchored at the inner end to a crossbeam, now is freely balanced on the bracket cluster, supporting purlins (horizontal timbers) at each end, thus giving the whole system something of the dynamic functionalism of High Gothic architecture. The interior is also much more elaborate. Richly detailed rounded vaults, or cupolas, are set in the ceiling over the principal images.

Practically nothing survives today of the Southern Sung capital of Hang-chou, described as the greatest city in the world by the Venetian traveler Marco Polo, who spent much of the time from 1276 to 1292 in the city. The dense population and confined space of Hang-chou forced buildings upward, and many dwellings were in three to five stories. While palace buildings in the southern part of the city were probably crowded together, temples and high-platformed viewing pavilions overlooking West Lake were buildings of fairylike beauty. They survive today only in the work of such Southern Sung landscape and architecture painters as Li Sung.

Variety of form, structural technique, detail, and decoration of Sung architecture reflect the sophistication of Sung culture and a new intellectual interest in the art. Master builders such as Yü Hao and the state architect Li Chieh were educated men. The latter is known today chiefly as the compiler of *Ying-tsao fa-shih* ("Building Standards"), which he presented to the throne in 1100. This illustrated work deals in encyclopaedic fashion with all branches of architecture: layout, construction, stonework, carpentry, bracketing, decoration, materials, and labour. The *Ying-tsao fa-shih* became a standard text, and, while it was influential in spreading the most advanced techniques of the time of its first publication in 1103, by codifying practice it may also have inhibited further development and contributed to the conservatism of later techniques.

In contrast with the greater uniformity of later periods, Sung architecture was experimental and increasingly diverse in nature. Two styles from the Southern Sung period can be inferred from early Japanese buildings. One style is called by the Japanese *Tenjiku-yo,* or "Indian style," but it actually originated on the southeastern Chinese coast, where tall stands of evergreens stood. It sometimes employed timber columns rising to about 20 metres, directly into which were inserted vertical tiers of up to 10 transverse bracket-arms. This stern and simple style is exemplified by the Great South Gate at Tōdai Temple, built in Nara, Japan, about 1180. Another style, dubbed *Kara-yo* ("T'ang"—*i.e.,* Chinese—style), was brought by Ch'an (Zen) Buddhist priests from the Hang-chou area and south to the new shogunal capital at Kamakura, where it can be seen in the 13th-century Reliquary Hall of Engaku Temple. It features unpainted wood siding with multilevel paneled walls (no plaster wall or lacquered columns) and much attention to elaborative detail. The effect is rich and dynamic and displays none of the simplicity one might expect of Ch'an architecture, so it is thought by some to represent more a Chinese regional style than anything specifically Ch'an.

Sculpture. The tradition of Buddhist image making continued through the Sung, Liao, and Chin dynasties with slowly diminishing force. The sculpture of southern Chinese temples in and around Hang-chou has been destroyed, but much survives in the north. A typical example is the lower temple of Hua-yen Temple at Ta-t'ung (1038), with its three central seated Buddha figures, 5 metres high, of lacquered clay, and its sensuous, exquisitely mannered bodhisattvas. A realistic strain in Sung, Liao, and Chin sculpture shares with Ch'an painting of the late Sung and Yüan dynasties the power to suggest spiritual concentration through the strongly marked features of the image. A similar realism was attained in lohan figures modeled in dry lacquer, a technique that was first used in China in the second half of the 6th century and of which the best surviving examples are to be found today at Nara in Japan. This realistic style of sculpture can also be seen at such sites as the Lung-men caves at Lo-yang, the Chin Tz'u at T'ai-yüan, and Ta-tsu in Szechwan.



Painted clay bodhisattva, 11th century, Liao dynasty. In the lower Hua-yen Temple, Ta-t'ung, Shansi province.
Gao Lishuang/ChinaStock Photo Library

Most majestic among Sung sculptures are images of Kuan-yin seated in a relaxed pose known as "royal ease." The body is sensuously modeled and the face full and slightly smiling; rich jewelry, scarves, and ribbons move over the surface, creating a dramatic play of light and shade that, like Baroque sculpture in Europe, was no doubt designed to win back adherents to the faith by its visual and emotional appeal. After the climax reached in this spectacular, somewhat florid style, there were no major developments in Chinese Buddhist sculpture, which preserved the essentials of this style, though with increasing coarseness, through the Yüan, Ming, and Ch'ing dynasties.

Painting and calligraphy. Settled conditions and a tolerant atmosphere helped to make the Northern Sung a period of great achievement in landscape painting. Li Ch'eng, a follower of Ching Hao who lived a few years into the Sung, was a scholar who defined the soft, billowing earthen formations of the northeastern Chinese terrain with "cloudlike" texture, interior layers of graded ink wash bounded by firmly brushed, scallop-edged contours. He is remembered especially for winter landscapes and for simple compositions in which he set a pair of tall, rugged, aging evergreens against a low, level view of desiccated landscape. As with Ching Hao and Kuan T'ung, probably none of his original work survives, but aspects of his style have been perpetuated in thousands of other artists' works. An even more formidable figure was the early

The painter Fan K'uan

11th-century painter Fan K'uan, who began by following Li Ch'eng's style, but turned to studying nature directly, and finally followed only his own inclinations. He lived as a recluse in the mountains of Shensi, and a Sung writer said that "his manners and appearance were stern and old-fashioned; he had a great love of wine and was devoted to the Tao." A tall landscape scroll, "Travelers Among Mountains and Streams" (National Palace Museum, Taipei), bearing his hidden signature, depicts peasants and pack mules emerging from thick woodland at the foot of a towering cliff that dwarfs them to insignificance. The composition is monumental, the detail is realistic, and the brushwork, featuring a stippling style known as "raindrop" strokes, is powerful and close-textured. While the details of the work are based on closely observed geographic reality (perhaps some specific site such as Mount Heng), a profoundly idealistic conception is revealed in the highly rational structure of the painting, which conforms closely to aspects of Taoist cosmology and numerology.

Other northern masters of the 11th century who helped to establish the great classical tradition were Hsü Tao-ning, Kao K'o-kung, and Yen Wen-kuei. The second half of the century was dominated by Kuo Hsi, who became an instructor in the painting division of the Imperial Hanlin Academy. His style combined the technique of Li Ch'eng with the monumentality of Fan K'uan; but it shows some advance, particularly in the effect of relief that he attained by shading with ink washes ("cloudlike" texture), a spectacular example of which is his "Early Spring" of 1072 (National Palace Museum, Taipei). He was a great decorator and liked to work on such large surfaces as plaster walls and standing screens. His observations on landscape painting were collected and published by his son Kuo Ssu under the title *Lin-ch'uan kao-chih* ("Lofty Record of Forests and Streams").

While the monumental realistic tradition was reaching its climax, quite another approach to painting was being expressed by a group of intellectuals that included the poet-statesman-artist Su Shih (or Su Tung-p'o), the landscape painter Mi Fei (Mi Fu), the bamboo painter Wen T'ung, the plum painter and priest Chung-jen Hua-kuang, and the figure and horse painter Li Kung-lin. Su and Mi, together with their friend Huang T'ing-chien, were also the foremost calligraphers of the dynasty, all three developing the tradition established by Chang Hsü, Yen Chen-ch'ing, and Huai-su in the mid-8th century. The aim of these artists was not to depict nature realistically—that could be left to the professionals—but to express themselves, to "satisfy the heart." They spoke of merely "borrowing" the forms of things in which for the moment to "lodge" their thoughts and feelings. In this amateur painting mode of the scholar-official (*shih-ta-fu hua,* later called *wen-jen hua*), skill was suspect because it was the attribute of the professional and court painter. The scholars valued spontaneity above all, even making a virtue of awkwardness as a sign of the painter's sincerity.

Mi Fu, an influential and demanding connoisseur, was the first major advocate and follower of Tung Yüan's boneless style, reducing it to mere ink dots (Mi *tien,* or "Mi dots"). This new technique influenced many painters, including Mi Fu's son Mi Yu-jen, who combined it with a subdued form of ink-splashing. Wen T'ung and Su Tung-p'o were both devoted to bamboo painting, an exacting art form very close in technique to calligraphy. Su Tung-p'o wrote poems on Wen T'ung's paintings, thus helping to establish the unity of the three arts of poetry, painting, and calligraphy that became a hallmark of the *wen-jen hua.* When Su Shih painted landscapes, Li Kung-lin sometimes put in the figures. Li was a master of *pai-miao* (plain line) painting, without colour, shading, or wash. He brought a scholar's refinement of taste to a tradition theretofore dominated by Wu Tao-tzu's dramatic style.

The emperor Hui-tsung as patron of the arts

The northern emperors were enthusiastic patrons of the arts. Hui-tsung, perhaps the most knowledgeable of all Chinese emperors about the arts, was himself an accomplished calligrapher (he developed a unique and extremely elegant style known as "slender gold") and a painter chiefly of birds and flowers in the realistic tradition stretching back to Huang Ch'üan and developed by subsequent court artists such as Ts'ui Po of the late 11th century. While meticulous in detail, his works were subjective in mood, following poetic themes that were calligraphically inscribed on the painting. A fine example of the kind of painting attributed to him is the minutely observed and carefully painted "Five-coloured Parakeet on Blossoming Apricot Tree" (Museum of Fine Arts, Boston).

Among the distinguished academicians at Hui-tsung's court were Chang Tse-tuan, whose extraordinarily realistic "Ching-ming Festival" scroll (Palace Museum, Peking) preserves a wealth of social and architectural information in compellingly artistic form, and Li T'ang, who fled to the south in 1127 and supervised the reestablishment of the northern artistic tradition at the new court in Hang-chou. Although Kuo Hsi's style remained popular in the north after the Chin occupation, Li T'ang's mature style came to dominate in the south. Li was a master in the Fan K'uan tradition, but he gradually reduced Fan's monumentality into more refined and delicate compositions and transformed Fan's small "raindrop" texture into a broader "axe-cut" texture stroke that subsequently remained a hallmark of most Chinese court academy landscape painting.

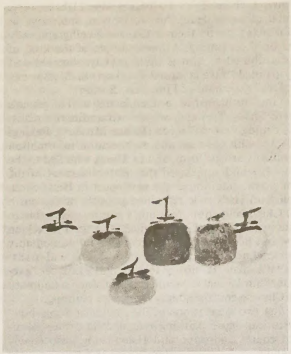
In the first two generations of the Southern Sung, however, historical figure painting regained its earlier dominance at court. Kao-tsung and Hsiao-tsung, respectively the son and grandson of the imprisoned Hui-tsung, sought to legitimize their necessary but technically unlawful assumption of power by supporting works illustrating the ancient classics and traditional virtues. Such works, by artists including Li T'ang and Ma Ho-chih, often include lengthy inscriptions purportedly executed by the emperors themselves. They represent the finest survival today of the ancient court tradition of propagandistic historical narrative painting in a Confucian political mode.

The Ma-Hsia school of painting

Subsequently, in the late 12th and early 13th centuries, the primacy of landscape painting was reasserted. The tradition of Li T'ang was turned, however, in an increasingly romantic and dreamlike direction by the great masters Ma Yüan, his son Ma Lin, Hsia Kuei, and Liu Sung-nien, all of whom served with distinction in the painting division of the Imperial Hanlin Academy. These artists used the Li T'ang technique, only more freely, developing the so-called "large axe-cut" texture stroke. Their compositions are often "one-cornered," depicting a foreground promontory with a fashionably rusticated building and a few stylish figures separated from the silhouettes of distant peaks by a vast and aesthetically poignant expanse of misty emptiness—a view these painters must have seen any summer evening as they gazed across Hang-chou's West Lake. The Ma family's works achieved a philosophically inspired sense of quietude, while Hsia Kuei's manner was strikingly dramatic in brushwork and composition. The Ma-Hsia school, as it came to be called, was greatly admired in Japan during the Muromachi and Momoyama periods, and its impact can still be found today in Japanese gardening traditions.

Toward the end of this period, Ch'an (Zen) Buddhist painting experienced a brief but remarkable florescence, stimulated by scholars abandoning the decaying political environment of the Southern Sung court for the monastic life practiced in the hill-temples across the lake from Hang-chou. The court painter Liang K'ai had been awarded the highest order, the Golden Girdle, between 1201 and 1204, but he put it aside, quit the court, and became a Ch'an recluse. What is thought to be his earlier work has the professional skill expected of a colleague of Ma Yüan, but his later paintings became freer and more spontaneous.

The greatest of the Ch'an painters was Mu-ch'i, or Fa-ch'ang, who reestablished the Liu-t'ung Monastery in the western hills of Hang-chou. The wide range of his work (which included Buddhist deities, landscapes, birds and animals, flowers and fruit) and the spontaneity of his style bear witness to the Ch'an philosophy that the "Buddha essence" is in all things equally and that only a spontaneous style can convey something of the sudden awareness that comes to the Ch'an adept in his moments of illumination. Perhaps his best-known work is his hastily sketched "Six Persimmons" (preserved and idolized in Japan), while a somewhat more conservative style is seen in his triptych



"Six Persimmons," hanging scroll attributed to Mu-ch'i (active mid-13th century), Southern Sung dynasty. Ink on paper. In the Daitoku Temple, Kyōto, Japan. Width 36.2 cm.
Daitoku-ji, Kyoto; photograph, Zen Cultural Laboratory

of three hanging scrolls with Kuan-yin flanked by a crane and gibbons (Daitoku Temple, Kyōto, Japan). Chinese connoisseurs disapproved of the rough brushwork and lack of literary content in Mu-ch'i's paintings, and none appears to have survived in China. His work, and that of other Ch'an artists such as Liang K'ai and Yü-chien, was collected and widely copied in Japan, however, forming the basis of the Japanese *sumi* tradition.

Ch'an Buddhism borrowed greatly from Taoism, both in philosophy and in painting manner. One of the last great Sung artists was Ch'en Jung, an official, poet, and Taoist who specialized in painting the dragon, a symbol both of the emperor and of the mysterious all-pervading force of the Tao. Ch'en Jung's paintings show these fabulous creatures emerging from amid rocks and clouds. They were painted in a variety of strange techniques, including rubbing the ink on with a cloth and spattering it, perhaps by blowing ink onto the painting.

High point in the history of Chinese pottery

Ceramics. The Sung dynasty marked a high point in the history of Chinese pottery, when technical mastery, refinement of feeling, and a natural spontaneity of technique were more perfectly balanced than at any time in Chinese history. The beauty of Sung wares, derived from the simplicity of the shapes and purity of glaze tone and colour, is not a lifeless perfection such as marks the palace wares of the Ch'ing dynasty; in Sung wares the touch of the potter's hand can still be perceived, and glazes have a depth and warmth that was later lost when a higher level of manufacturing skill was attained.

It is convenient to group Sung wares geographically: the chief northern wares are Ting, Ju, Chün, northern celadon, Tz'u-chou, and brown and black glazed wares; those of southern China include Ching-te-chen whiteware (*ying-ch'ing,* or *ch'ing-pai*), Chi-chou wares, celadons, and black wares of Fukien. However, many new kiln sites have been located, and it is now known that one kiln often produced several quite different wares, and decorated stonewares named from the principal factory at Tz'u-chou in southern Hopeh were made in many kilns across the breadth of northern China.

White porcelain made at Chien-tz'u-ts'un in south-central Hopeh was already being produced for the northern courts in the Five Dynasties and continued as an Imperial ware to the beginning of the 12th century. Very finely potted and sometimes decorated with freely incised plants, fish, and birds under the glaze or later with mold-made designs in relief, this Ting ware is directly descended from the northern whitewares of the T'ang dynasty. Supposedly because of Hui-tsung's dissatisfaction with Ting ware, it was replaced in the late Northern Sung by another official ware known as Ju, the rarest and most highly prized of all Chinese ceramics (until recently, only some 60 examples were known). Representing Hui-tsung's celebrated aestheticism, the low-fired Ju stoneware is distinguished by a seemingly soft, milky glaze of pale blue or grayish green with hair-thin crackle. The glaze covers a pale gray or buff body that is usually simple in shape yet highly sophisticated and exquisitely tasteful in effect. Ju ware was produced for only a few years before Hui-tsung's sudden demise. The Ju kilns defied identification until 1986, when they, along with the remains of a workshop, were located at Ch'ing-liang-ssu, more than 160 kilometres southwest of the capital. Another 37 intact examples were soon afterward excavated there. Typical of other kilns, the Ju kilns varied their productions, turning out Tz'u-chou stoneware and Yao-chou-type celadons like those discovered at Yao-an, north of Sian.

Chün stoneware

A sturdy stoneware covered with a thick lavender-blue glaze was made at Chün-chou in Honan. This Chün ware is sometimes marked with splashes of purple or crimson produced by copper oxide. On the finest Chün wares, which are close to Ju in quality, these splashes are used with restraint, but on later Chün-type wares manufactured at Ching-te-chen and near Canton too much purple often gives vessels or flowerpots a mottled, lurid hue that Ming connoisseurs were wont to label "mule's liver" or "horse's lung."

Somewhat related to Chün wares are sturdily potted jars, vases, and bowls with lustrous black or brown glazes. Those that are decorated with flowers and leaves painted in an oxidized rust brown constitute an enormous family of Tz'u-chou wares made for domestic and funerary use in numerous northern China kilns, and they are still being produced in some factories today. Tz'u-chou techniques of decoration included free brush painting under the glaze, carving or scratching (sgraffito work) through one slip to another of a different colour, and painting over the glaze in low-fired colours. The earliest known example of overglaze painting in the history of Chinese pottery bears a date equivalent to 1201. In both the variety and the vigour of their forms and decoration, Tz'u-chou stonewares present a strong contrast to the restraint and exquisite taste of the courtly wares.



Tz'u-chou stoneware vase probably from Chü-lu-hsien, Hopeh province, early 12th century. Northern Sung or Chin dynasty. Dragon decor with white and engraved black slip. In The Nelson-Atkins Museum of Art, Kansas City, Mo., U.S. Height 56.7 cm.
The Nelson-Atkins Museum of Art, Kansas City, Missouri; purchase: Nelson Trust (36-116)

South China wares

After the Sung capital was reestablished at Hang-chou, the finest wares obtainable were once more supplied to the court. These southern *kuan* wares were made for a short time in kilns close to the palace under the direction of the Office of Works. Later, the kilns were established near the altar for sacrifices to heaven and earth, Chiao-t'an, outside the south gate of the city. Chiao-t'an *kuan* ware had a dark opaque body and a beautiful bluish gray layered glaze. A deliberately formed crackle, caused by the shrinking of the glaze as the vessel cooled after firing, is the only ornament on this exquisite ware.

The southern *kuan* was the finest of a huge family of celadons produced in an increasing number of kilns in southeastern China. Lung-ch'üan in southern Chekiang made a fine celadon with bluish green glaze, the best of which was almost certainly supplied to the court and may hence be classed as *kuan.* A variant with strongly marked crackle became known as *ko* ware in deference to the tradition that it was made by the elder brother (*ko*) of the director of the Lung-ch'üan factory. Among the wide range of shapes made in Sung celadon are those derived from forms of archaic bronzes, such as *li, ting,* and *tsun,* testifying to the increasingly antiquarian taste of court and gentry.

Meanwhile, a small factory at Ching-te-chen in Kiangsi was growing to meet the vast increase in the population of southern China. In the Sung, its most characteristic ware was a fine, white, sugary porcelain covered with a transparent, slightly bluish glaze; the ware has been known since Sung times as *ch'ing-pai* ("bluish white") but is called by modern Chinese dealers *ying-ch'ing* ("shadowy blue"). *Ch'ing-pai* ware is very thinly potted, the decoration carved in the clay body or applied in raised slip or beading under the glaze. Sung *ch'ing-pai* wares are the predecessors of a vast output of fine white Ching-te-chen porcelain that was to dominate the Chinese pottery industry during the Yüan, Ming, and Ch'ing dynasties. Other whitewares were made at Yung-ho near Chi-an in Kiangsi.

Kilns in the wooded hills around Chien-yang in northern Fukien produced almost nothing but heavily potted stoneware teabowls covered with a thick black glaze. The finest and rarest of these Chien ware bowls have streaky "hare's fur" or iridescent "oil spot" effects that were much prized by Japanese tea masters, who called this ware *temmoku* after T'ien-mu, the sacred Buddhist mountain in Chekiang province that was near the port from which the ware was shipped to Japan. Yung-ho kilns also turned out a coarse variety of *temmoku* and experimented with novel decorative effects produced by laying floral paper cuts or skeleton leaves under the glaze before firing.

Textiles. Lacking a Sung Shōsō-in repository for decorative arts, knowledge of the textiles of this period is even sketchier than knowledge of those of the T'ang. The

K'o-ssu silk tapestry

main Sung achievement was the perfecting of *k'o-ssu,* an extremely fine silk tapestry woven on a small loom with a needle as a shuttle. The technique appears to have been invented by the Sogdians in Central Asia, improved by the Uighurs, and adapted by the Chinese in the 11th century. The term *k'o-ssu* (literally "cut silk") derives from vertical gaps between areas of colours, caused by the weft threads not running right across the width; it has also been suggested that the word is a corruption of the Persian *qazz* or Arabic *khazz,* referring to silk and silk products. *K'o-ssu* was used for robes, silk panels, and scroll covers and for translating painting into tapestry. In the Yüan dynasty, panels of *k'o-ssu* were exported to Europe, where they were incorporated into cathedral vestments.

Yüan dynasty (1206–1368). Although the Mongol conquest made China part of an empire that stretched from Korea to Hungary and opened its doors to foreign contacts as never before, this short-lived dynasty was oppressive and corrupt. Its later decades were marked by social and administrative chaos in which the arts received little official encouragement. The Mongols distrusted the Chinese intelligentsia, relying primarily on Central Asians for government administrative functions. Nevertheless, some influential Chinese writers recognized that the Mongols brought a sense of martial discipline that was lacking in the Sung, and after 1286 an increasing number of Chinese scholars were persuaded to enter government service, undoubtedly hoping to influence their rulers to adopt a more benign policy toward the Chinese people.

Architecture. Little remains of Yüan architecture today. The great palace of Kublai Khan in the Yüan capital Ta-tu ("Great Capital"; now Peking) was entirely rebuilt in the Ming dynasty. Excavations demonstrate that the Yüan city plan is largely retained in that of the Ming; originally conceived under the combined influence of Liu Ping-chung and non-Chinese Muslims such as Yeheidie'er, it appears to be thoroughly Chinese in concept. More detailed information survives only in first-generation Ming dynasty court records and in the somewhat exaggerated description of Marco Polo. This architecture was probably little advanced in point of building technique over those of the Liao and Chin palaces on which they were modeled. The ornate features of their roofs, their bracketing systems, the elevated terraces, and the tight juxtaposition of the buildings are reflected in architectural paintings of the period by such artists as Wang Chen-p'eng, Hsia Yung, and Li Jung-chin. Perhaps the only original Yüan buildings in Peking today are the Drum Tower to the north of the city and the White Pagoda built by Kublai in the stupa form most commonly seen today in the Tibetan *chorten.* The Mongols were ardent converts to Tibetan Buddhism and tolerant of the Taoists, but they seem to have found existing temples enough for their purposes, for they made few new foundations.

Racho/Photo Researchers



Detail of the marble relief of a guardian king on Chü-yung-kuan Gate, Peking, 1345, Yüan dynasty.

Florescence of Buddhist sculpture

Sculpture. Buddhist sculpture was one of the few arts that flourished under the patronage of the Mongols, who inherited craftsmen and techniques from the Chin. Yüan style is thus a continuation of the northern tradition, marked by a rather more emphatic modeling and greater richness of detail. A richly sculptured marble gate, constructed in 1342–45 at the Chü-yung Pass between Peking and the Great Wall to the north, boldly illustrates in shallow relief a variety of Buddhist and Taoist divinities and guardian figures; it represents perhaps the finest sculptural work of this or any later period. At this time esoteric Tantric sects were popular, and their complex iconography demanded images with multiple arms and attributes. Some especially fine, realistic Buddhist bronzes may be attributed to a special government department responsible for image casting by the lost-wax process. The dry lacquer technique, invented before the T'ang dynasty, was



"Twin Pines and Level View," detail of right half of a hand scroll by Chao Meng-fu (1254–1322), Yüan dynasty. Ink on paper. In The Metropolitan Museum of Art, New York City. Size of entire hand scroll 28.7 cm × 1.07 m.

The Metropolitan Museum of Art, New York City, gift of The Dillon Fund, 1973 (1973.120.5); all rights reserved, The Metropolitan Museum of Art

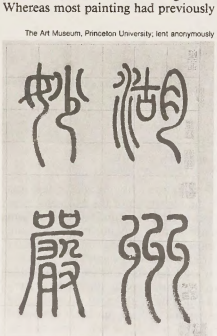
popularized in the Yüan dynasty by the master craftsman Liu Yüan, a scholar and Taoist who made a deep study of Indian iconography before executing the deities for a temple commissioned by the emperor in 1270.

Painting and calligraphy. One school that flourished under Yüan official patronage was that of Buddhist and Taoist painting; important wall paintings were executed at the Yung-lo Temple in Shansi (now restored and moved to Jui-ch'eng). A number of royal patrons, including Kublai, the emperors Buyantu and Tog-temür, and Kublai's great-granddaughter Sengge, built an Imperial collection of important early works and also sponsored paintings that emphasized such themes as architecture and horses. Still, their activities were not a match for Sung royal patronage, and it was in this period that the amateur art of painters of the scholar class (in the tradition of Su Shih and his late Northern Sung colleagues) first came to dominate Chinese painting standards.

The restriction of the scholars' opportunities at court and the choice of many of them to withdraw into seclusion rather than serve the Mongols created a heightened sense of class identity and individual purpose, which in turn inspired their art. Eremitic rather than courtly values now shaped the art of painting as never before, and a stylistic gulf sprang up between literati painters and court professionals that was not bridged until the 18th century. Whereas most painting had previously displayed technical refinement and had conservatively transmitted the heritage of the immediate past, gradually evolving through modest individual departures, the literati thenceforth typically based their styles on a wide-ranging knowledge of distant stylistic precedents, selectively chosen and radically transformed by means of expressive calligraphic brushwork. Style and subject were both intended to reflect closely the artist's own personality and mood rather than conforming to the wishes of a patron. Typical were the simply brushed orchid paintings of Cheng Ssu-hsiao, who painted this traditional symbol of political loyalty without any ground beneath as a comment on the grievous loss of China to foreign domination.

Ch'ien Hsüan was among the first to define this new direction. From Wu-hsing in Chekiang, he steadfastly declined an invitation to serve at court, as reflected in his painting style and themes. A conservative painter before the Mongol conquest, especially of realistic flowers and birds, he altered his style to incorporate the primitive qualities of ancient painting, favouring the T'ang blue-and-green manner in his landscape painting, stiff or peculiarly mannered renditions of vegetation and small animals, and the archaic flavour of clerical script in his brushwork. Calligraphy became a part of his design and frequently confirmed through historical references a link between subject matter and his eremitic choice of lifestyle. Like many Chinese scholars who espoused this amateur ideal, Ch'ien Hsüan was obliged by demeaning circumstances to exchange his paintings in return for his family's livelihood.

The Art Museum, Princeton University; lent anonymously



"Record of the Miao-yen Temple in Hu-chou," seal script by Chao Meng-fu, c. 1309–10, Yüan dynasty. Detail of frontispiece of a hand scroll, ink on paper. In The Art Museum, Princeton University, New Jersey, U.S. Height 35.2 cm.

The painter and calligrapher Chao Meng-fu

The most distinguished of the scholar-painters was Chao Meng-fu, a fellow townsman and younger follower of Ch'ien Hsüan who became a high official and president of the Imperial Hanlin Academy. In his official travels he collected paintings by Northern Sung masters that inspired him to revive and reinterpret the classical styles in his own fashion. A notable example is "Autumn Colours in the Ch'iao and Hua Mountains" (1296; National Palace Museum, Taipei), a nostalgic, deliberately archaistic landscape in the T'ang manner. The hand scrolls "Twin Pines and Level View" (The Metropolitan Museum of Art, New York City) and "Water Village" (1302; Palace Museum, Peking) exemplify his reinterpretation of past masters (Li Ch'eng and Tung Yüan, respectively) and furthered the new direction of scholarly landscape painting by applying the standards and techniques of calligraphy to painting.

The Yüan produced many fine calligraphers, including Chao Meng-fu, who was the most influential, Yang Wei-chen, and Chang Yü. The period was less innovative in calligraphy than in painting, however, and Chao's primary accomplishment was to sum up and resynthesize the past. His well-studied writing style was praised in his time for its breadth of historical understanding, and his standard script became the national model for book printing, but he was later criticized for a lack of daring or expression of personality, for a brush style too sweet and pleasing.

Other gentlemen-painters who worked at the Yüan court

perpetuated more conservative Sung styles, often rivaling or even surpassing their Sung predecessors in the process. Jen Jen-fa worked in great detail and was perhaps the last of China's great horse painters; he defended his court service through both the style and theme of his paintings. Li K'an carefully studied the varieties of bamboo during his official travels and wrote a systematic treatise on painting them; he remains unsurpassed as a skilled bamboo painter. Kao K'o-kung followed Mi Fu and Mi Yu-jen in painting cloudy landscapes that symbolized good government. Wang Mien, who served not the Mongols but anti-Mongol forces at the end of the dynasty, set the highest standard for the painting of plums, a symbol of irrepressible purity and, potentially, of revolutionary zeal.

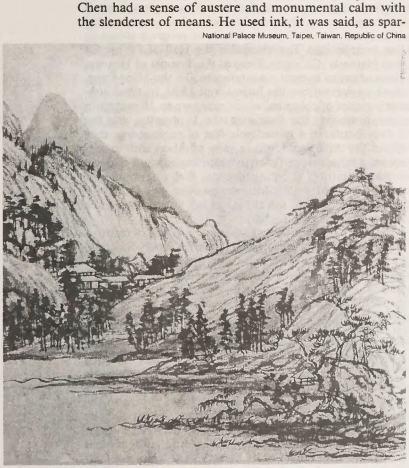
Ideals of the retired scholars

In retrospect, however, it was the ideals of the retired scholars that had the most lasting effect on later Chinese art. This may be summed up as individuality of expression, brushwork more revealing of the inner spirit of the subject—or of the artist himself—than of outward appearance, and suppression of the realistic and decorative in favour of an intentional plainness, understatement (*p'ing-tan*), and awkwardness (*cho*), which marks the integrity of the gentleman suspicious of too much skill. Four masters of the middle and later Yüan, all greatly influenced by Chao Meng-fu, came to be regarded as the foremost exponents of this philosophy of painting in the Yüan period. Huang Kung-wang, a Taoist recluse, was the oldest. His most revered and perhaps only authentic surviving work is the hand scroll "Dwelling in the Fu-ch'un Mountains" (National Palace Museum, Taipei), painted with dynamic brushwork during occasional moods of inspiration between 1347 and 1350. Wu Chen was a poor Taoist diviner, poet, and painter who, like Huang Kung-wang, was inspired by Tung Yüan and Chü-jan, whose manner he rendered, in landscapes and bamboo painting alike, with blunt brushwork, minimal motion, and utmost calm.

The third of the Four Masters of the Yüan dynasty was Ni Tsan, a prosperous gentleman and bibliophile forced by crippling taxation to give up his estates and become a wanderer. As a landscapist he eliminated all depictions of human beings. He thus reduced the compositional pattern of Li Ch'eng (symbolizing lofty gentlemen in isolation from the court) to its simplest terms, achieving as Wu Chen had a sense of austere and monumental calm with the slenderest of means. He used ink, it was said, as sparingly as if it were gold. Quite different was the technique of the fourth Yüan master, Wang Meng, a grandson of Chao Meng-fu. His brushwork was dense and energetic, derived from Tung Yüan but tangled and hoary and thereby imbued with a feeling of great antiquity. He often drew heavily from Kuo Hsi or from what he perceived as T'ang traditions in his landscape compositions, which he filled with scholarly retreats. He sometimes used strong colours as well, which added a degree of visual charm and nostalgia to his painting that was lacking in the other three masters' work.



"The Jung-hsi Studio," hanging scroll by Ni Tsan, 1372, late Yüan dynasty. Ink on paper. In the National Palace Museum, Taipei, Taiwan. 74.7 × 35.5 cm.
National Palace Museum, Taipei, Taiwan, Republic of China



"Dwelling in the Fu-ch'un Mountains," detail from a hand scroll by Huang Kung-wang, dated 1350, Yüan dynasty. Ink on paper. In the National Palace Museum, Taipei. Complete scroll 59.9 m × 33 cm.
National Palace Museum, Taipei, Taiwan, Republic of China

Models for later scholar-painters

The combination in the Four Masters of a consistent philosophical and political attitude and a wide range of ink techniques made them models for later scholar-painters, both in their lives and in their art. It is impossible to appreciate the work of the landscape painters of the Ming and Ch'ing dynasties unless one is aware of how acutely conscious they were of their debt to the Yüan masters and how frequently they paid tribute to them both in their style and in their inscriptions. Thenceforward, indeed, the artist's own inscription, as well as the colophons of admirers and connoisseurs, became an integral part of the total work of art.

Ceramics. It is a paradox that the Mongol occupation, while it destroyed much, also shook China free from the static traditions and techniques of the late Southern Sung and made possible many innovations, both in painting and in the decorative arts. The north was least progressive, and the main centre of pottery activity shifted permanently to the south. The northern traditions of Chün and Tz'u-chou ware continued through the Chin and Yüan, bolder but coarser than before. New shapes included a heavy, wide-mouthed jar, sometimes with decoration boldly carved through a black or brown slip or painted in two or three

colours. These new techniques and the overglaze painting already developed in the Chin dynasty prepared the way for the three- and five-colour wares of the Ming.

While no Yüan celadon has the perfection of colour of Sung *kuan* and Lung-ch'üan wares, being more olive-green in tone, the quality is high. And the variety of decorative techniques is far wider than that of the Sung. These include raised relief designs molded under the glaze, fish and dragons raised "in the biscuit" (that is, unglazed) in relief, and iron-brown spots that the Japanese call *tobi seiji* ("flying celadon"). Vases and dishes were now sturdily potted in porcelain, often mold-made, and of considerable size.

Factories at Ching-te-chen were expanding rapidly. While their products included celadon, their chief output, as before, was white porcelain, including richly modeled figurines of Kuan-yin and other Buddhist deities. *Ch'ing-pai* was now decorated with floral motifs and beading in raised relief or incised under the glaze, the most elaborate pieces combining flowers and vines in appliqué relief with openwork panels. A stronger, less sugar-white porcelain with molded or incised decoration was produced—called *shu-fu* ware from the presence of these characters (perhaps referring to a central government agency) under the thick, unctuous glaze.

Use of cobalt blue

The earliest evidence of the use of cobalt blue, probably imported from the Middle East, is seen in its application as an underglaze pigment on fragments dating to the late 8th or early 9th century that were unearthed at Yang-chou in 1983. The occasional use of underglazed cobalt continued in the Northern Sung. It was not until the Yüan dynasty, however, that underglazed blue decoration began a rapid rise in popularity. It was applied on fine white porcelains of the *shu-fu* type and combined with Islämic decorative taste. These blue-and-white wares soon became the most popular of all Chinese ceramics, both at home and abroad. A pair of richly ornate temple vases dated 1351 (in the Percival David Foundation in London) are proof that the technique was fully mastered by that time. The finest Ching-te-chen examples were reserved for the court, but coarse varieties were made in southern China for trade with Southeast Asia or for export to the Middle East. Experiments also were made with painting in underglaze copper red, but it was difficult to control and soon abandoned.

Lacquerwork. While lacquer continued to be made in bolder versions of the undecorated T'ang and Sung shapes, notable advances included incising and engraving and filling the lines with gold leaf or silver powder. The most important innovation was the carving of pictorial designs, floral patterns, or dragons through a thick coating of red or, less frequently, black lacquer. A connoisseur's manual, *Ko ku yao lun* ("Essential Criteria of Antiquities") by Ts'ao Chao, says that at the end of the Yüan dynasty Chang Ch'eng and Yang Mao, pupils of Yang Hui, were noted for this technique.

Ming dynasty (1368–1644). The restoration of a native dynasty made China once more a great power. Ming felt a kinship with the heyday of the T'ang dynasty that is reflected in the vigour and rich colour of Ming arts and crafts. Early in the 1400s, China again expanded into Central Asia, and maritime expeditions brought Central Asian products around the Indian Ocean to its own shores. Chinese pottery exports also greatly increased. The 15th century was a period of settled prosperity and great achievement in the arts, but the last century of the dynasty was marked by corruption at court and deep discontent among the scholar-gentry that is reflected in their painting.

Architecture. The first Ming emperor established his capital at Nanking (Nan-ching, "Southern Capital"), surrounding it with a wall more than 30 kilometres in length, one of the longest in the world. The palace he constructed no longer exists. In 1402, a son of the founding Ming emperor enfeoffed at the old Yüan capital usurped the throne from his nephew, the second Ming ruler, and installed himself as the Yung-lo Emperor. He rebuilt the destroyed Mongol palaces and moved the Ming capital there in 1421, renaming the city Peking (Pei-ching, "Northern Capital"). His central palace cluster, the Forbidden City, is the foremost surviving Chinese palace compound, maintained and successively rebuilt over the centuries. In a strict hierarchical sequence, the palaces lie centred within the bureaucratic auspices of the Imperial City, which is surrounded by the metropolis that came to be called the "inner city," in contrast to the newer (1556) walled southern suburbs, or "outer city." A series of eight major state temples lay on the periphery in balanced symmetry, including temples to the sun (east) and moon (west) and, to the far south of the city, the huge matched temple grounds of heaven (east) and agriculture (west). Close adherence to the traditional principles of north-south axiality, walled enclosures and restrictive gateways, systematic compounding of courtyard structures, regimentation of scale, and a hierarchy of roofing types were all intended to satisfy classical architectural norms, displaying visually the renewed might of native rulers and their restoration of traditional order.

The Forbidden City

Central to this entire arrangement are three great halls of state, situated on a high, triple-level marble platform (the number three, here and elsewhere, symbolic of heaven and of the Imperial role as chief communicant between heaven and earth). The southernmost of these is the largest wooden building in China (roughly 65 by 35 metres), known as the Hall of Supreme Harmony. (The names and specific functions of many of the main halls were changed several times in the Ming and Ch'ing.) To their north lies a smaller-scale trio, the main halls of the Inner Court, in which the emperor and his ladies resided. The entire complex now comprises some 9,000 rooms (of an intended 9,999, representing a perfect yang number). The grandeur of this palatial scheme was matched by the layout of a vast Imperial burial ground on the southern slopes of the mountain range to the north of Peking, not far from the Great Wall, which eventually came to house 13 royal mausoleums, with an elaborate "spirit way" and accompanying ritual temple complexes.

In its colossal scale, the monumental sweep of its golden-tiled roofs, and its axial symmetry, the heart of the Forbidden City is unsurpassed as a symbol of Imperial power. In architectural technique, however, the buildings represent a decline from the adventurous planning and construction of the Sung period. Now the unit is a simple square or rectangular pavilion with few projections or none, and the bracketing system is reduced to a decorative frieze with little or no structural function. Instead, emphasis is placed upon carved balustrades, rich colour, and painted architectural detail. This same lack of progress shows in Ming temples also. Exceptional is the Hall of Prayer for Good Harvests (Ch'i-nien Tien) at the Temple of Heaven, a descendant of ancient state temples like the Ming-t'ang. It took its present circular form about 1530. Its three concentric circles of columns, which range up to 18 metres in height, symbolize the 4 seasons, the 12 months, and the 12 daily hours; in a remarkable feat of engineering, they support the three roof levels (a yang number) and, in succession, a huge square brace representing earth, a massive circular architrave denoting heaven, and a vast interior cupola decorated with golden dragons among clouds. (In its final rebuilding, in the 1890s, its tall timbers had to be imported from Oregon, U.S.)

Painting. The first Ming emperor was a highly distrustful personality whose vengeful focus fell upon Su-chou, the local base of his chief rival for the throne as well as home to the Yüan period literati painting movement. So many artists became victim to his recriminations, typically for political rather than artistic reasons, that this novel movement in Chinese painting history was nearly halted. Among those literati painters who lost their lives in this manner were Wang Meng, Chao Yüan, Hsü Pen, Ch'en Ju-yen, Chang Yü, Chou Wei, and Sheng Chu. Rejecting the individualist standard of literati painting, early Ming emperors who revived the custom of summoning painters to court sought instead to create a cultural bridge to the last native regimes, the Tang and Sung. Although they revived Sung professional court styles, they never organized their painters into a central teaching academy and indeed sometimes dealt quite harshly with them. Scholar-painters, increasingly few in number in the early Ming, stayed at home in the south, further widening the gulf between themselves and court artists.

Court painters

Early Ming court painters such as Hsieh Huan and Li Tsai at first revived the T'ang blue-and-green and Northern Sung court styles of Kuo Hsi. Pien Wen-chin and his follower Lü Chi carried forward the bird and flower painting tradition of Huang Ch'üan, Ts'ui Po, and the Sung emperor Hui-tsung. Gradually, however, the Southern Sung styles of the landscape artists Li T'ang, Ma Yüan, and Hsia Kuei came to hold sway, beginning with Tai Chin, who served under the fifth emperor, the Hsüan-te Emperor (himself a painter of moderate ability). Nevertheless, Tai Chin, who was opposed in the Peking capital by jealous court rivals and who found the restrictions there intolerable (as did many others who followed), was affected by the calligraphically inspired scholars' art: his brushwork shows far greater freedom than is found in his Southern Sung models.

Like Tai Chin, many professional painters came to Peking from the old Southern Sung capital region around Hangchou, and they were said to belong to the Che school of painting. Many of the so-called Che school artists were in fact scholars disgruntled with the autocratic Ming politics and drawn to Taoist eremitic themes and eccentric brushwork. Most dazzling among them, perhaps, was Wu Wei, whose drunken bouts at court were forgiven out of admiration for his genius with the brush.

Among the few important amateur painters to hold a scholarly position at the early Ming court was Wang Fu, who survived a long period of banishment to the frontier under the first emperor to return as a court calligrapher. He became a key figure in the survival and transmission of Yüan literati style and the first to single out the masters Huang Kung-wang, Wu Chen, Ni Tsan, and Wang Meng as models. Other early Ming scholar-official painters in the Yüan tradition were the bamboo painter Hsia Ch'ang and Liu Chüeh.

The Wu school of landscape painting

The Wu district of Kiangsu, in which Su-chou lies, gave its name to the Wu school of landscape painting, dominated in the late 15th century by Liu Chüeh's friend and pupil Shen Chou. Shen Chou never became an official but devoted his life to painting and poetry. He often painted in the manner of the Yüan masters, but his interpretations of Ni Tsan and Wu Chen are more clearly structured, firmer in brushwork, and unsurpassed in all Chinese art for their humane feeling, while the gentle and unpretentious figures he introduced give his paintings great appeal. Shen Chou commanded a wide range of styles and techniques, on which he impressed his warm and vigorous personality. He also became the first to establish among the literati painters a flower painting tradition. These works, executed in the "boneless" fashion developed by 10th-century court artists but with the freedom of such late Sung Ch'an painters as Mu-ch'i, were followed with greater technical versatility by Ch'en Shun and Hsü Wei in the late Ming and then by Shih-t'ao (Tao-chi) and Chu Ta of the early Ch'ing. Their work, in turn, served as the basis for the late-19th–20th-century revival of flower painting.

Shen Chou's younger contemporary and friend Wen Cheng-ming showed an even greater interest in the styles of the past, which he reinterpreted with a refined and scholarly precision. He, too, had many styles and was a distinguished calligrapher. He was an active teacher of painting as well, and among his more gifted pupils were his son Wen Chia and his nephew Wen Po-jen. Their landscapes display a lyrical delicacy in composition, touch, and colour, qualities that in lesser late Ming artists of the Wu school degenerated into a precious and artificial style.

Three early 16th-century professional Su-chou masters, Chou Ch'en, Ch'iu Ying, and T'ang Yin, established a standard somewhat different from that of the scholarly Wu group, never renouncing the professional's technical skills yet mastering the literary technique as well. They achieved a wide range, and sometimes a blend, of styles that could hardly be dismissed by scholarly critics and won great popular acclaim. In the succeeding generations, other painting masters similarly helped confuse the distinction between amateur and professional standards, and, in the early 17th century, a number of these also showed the first influence of the European technique that had been brought to China through engravings and then oil paintings by Matteo Ricci and other Jesuit missionaries after 1600. Among these painters were the landscapists Wu Pin from Nanking, Chang Hung from Su-chou, and Lan Ying from Ch'ien-t'ang in Chekiang province. Two who initiated the first major revival of figure painting since Sung times, possibly resulting from the encounter with Western art, were the southern painter Ch'en Hung-shou and the Peking artist Ts'ui Tzu-chung. Perspective and shading effects appear among other naturalistic features in the art of this generation, along with a newfound interest in saturated colours and an attraction to formal distortion, which may have derived in part from a fascination with the unfamiliar in Western art. Beyond the revived interest in naturalism, which seems to have inspired in some artists a renewed attention to Five Dynasties and Sung painting (as the last period in which Chinese artists had displayed knowledge about such matters), there occurred an even more fundamental questioning of contemporary standards and a deep disillusionment that were part of the spirit of this period of national decline. The breakdown of orthodoxy reached an extreme form in Hsü Wei. In his explosive paintings, chiefly of flowers, plants, and bamboo, he showed an absolute mastery of brush and ink and a total disregard of tradition.

Standing above all others of this period in terms of his-

The Nelson-Atkins Museum of Art, Kansas City, Missouri, purchase: Nelson Trust (46-51/2)



"Poet on a Mountain Top," album leaf mounted as a hand scroll by Shen Chou (1427–1509). Ming dynasty. Ink on paper or ink and light colour on paper. In The Nelson-Atkins Museum of Art, Kansas City, Mo., U.S. 38.7 × 60.2 cm.

Influence of Tung Ch'i-ch'ang

torical impact, the theorist, critic, and painter Tung Ch'i-ch'ang saw the proliferation of styles as a symptom of the decline in morale of the scholar class as the Ming became increasingly corrupt. His aim to reestablish standards in landscape painting paralleled a movement to restore traditional virtue to government. In his brief but influential essay *Hua shuo* ("Comments on Painting"), he set out what he held to be the proper lineage of scholarly painting models, from Wang Wei of the T'ang through Tung Yüan and Chü-jan of the Five Dynasties, Su Shih and Mi Fu of the Sung, and Huang Kung-wang, Wu Chen, Ni Tsan, and Wang Meng of the Yüan to Shen Chou and Wen Cheng-ming. He labeled these artists as "Southern school" in reference to the Southern school of Ch'an Buddhism and its philosophy of spontaneous enlightenment, while he rejected such "Northern school" (*i.e.,* gradualist, pedantic) artists as Kuo Hsi, Ma Yüan, Hsia Kuei, and Ch'iu Ying. Tung believed that the great painting models were highly creative individuals who, to be followed effectively, had to be creatively reinterpreted. Appropriately, his own landscape painting was often quite original, sometimes daringly so, even while based on a systematic reduction and synthetic reintegration of past styles. However, having breathed new life into a troubled tradition by looking inward and to the past, his reinterpretations (particularly of the styles of Tung Yüan and Chü-jan) set an ideal beyond which his contemporaries and followers could not go without either a great leap of imagination, a direct return to nature, or a departure from the historical core of Chinese painting standards. Only a few, in the early Ch'ing, could achieve this.

Woodblock printing

One further feature of late Ming art was the florescence of woodblock printing, including the appearance of a sophisticated tradition of polychrome printing, done in imitation of painting. Among the earliest major examples were the *Fang-shih mo p'u* of 1588 and *Ch'eng-shih mo yüan* of 1606 ("Mr. Fang Yü-lu's Ink Catalog" and "Mr. Ch'eng Ta-yüeh's Ink Garden," respectively), which utilized graphic designs by significant artists to promote the products of Anhwei province's foremost manufacturers of ink sticks. The *Shih-chu-chai shu-hua-p'u* ("Ten Bamboo Studio Manual of Painting and Calligraphy"), produced by Hu Cheng-yen between 1619 and 1633, set the highest standard for polychrome woodblock printing and helped influence the development of colour printing in Japan. Painters such as Ch'en Hung-shou participated in print production in forms ranging from book illustration to playing cards, while others, including Hsiao Yün-ts'ung, generated high-quality topographical illustrations. Through such artists, the medium came to influence painting as well as be influenced by it.

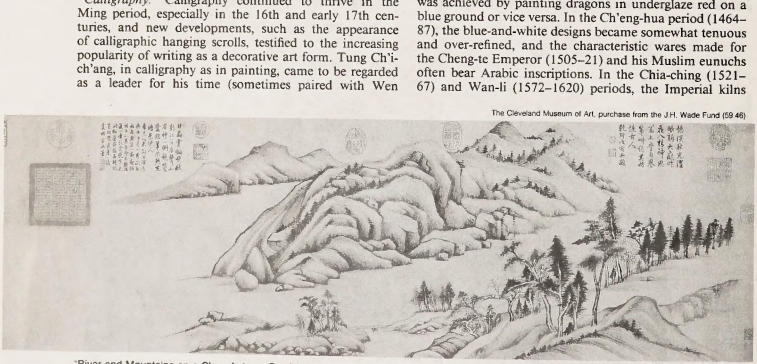
Calligraphy. Calligraphy continued to thrive in the Ming period, especially in the 16th and early 17th centuries, and new developments, such as the appearance of calligraphic hanging scrolls, testified to the increasing popularity of writing as a decorative art form. Tung Ch'i-ch'ang, in calligraphy as in painting, came to be regarded as a leader for his time (sometimes paired with Wen Cheng-ming, of the previous generation). He developed a style as historically conscious as, yet somewhat more individualized than, that of Chao Meng-fu of the Yüan, the previous acknowledged master. But Tung, like Chao, was subjected to criticism for too attractive a style. Furthermore, he was surrounded by numerous middle to late Ming masters who greatly enriched the art of the period with strikingly personal styles (especially in the execution of cursive and semicursive scripts) that showed the influence of the middle T'ang and Northern Sung dynasty individualists. They included Ch'en Hsien-chang, Chu Yün-ming, Ch'en Shun, Wang Wen, Hsü Wei, Chang Jui-t'u, Ni Yüan-lu, Huang Tao-chou, Shih K'o-fa, and Wang To.

Sculpture. There was a great deal of rebuilding and refurbishing of temples in this period of national recovery, but in point of style Ming religious sculpture represents no significant development except in grander scale, greater realism, and richer colours. Colossal figures lining the approaches to early Ming Imperial tombs at Nanking reflect a deliberate revival of T'ang style. More typically Ming are such architectural details as glazed pottery figures decorating roof ridges, whose vigorous modeling is derived from traditions of popular art. The best Ming sculpture is revealed in small ornamental pieces: animals and figures carved in jade, ivory, wood, and other materials, some of the most beautiful of which are Taoist and Buddhist figurines made in ivory-white porcelain at Te-hua in Fukien.

Expansion of pottery production

Ceramics. While northern traditions of Tz'u-chou and Chün ware continued to decline, pottery production in the south expanded. It was chiefly centred on Ching-te-chen, an ideal site because of the abundance of minerals used for porcelain manufacture—china clay and china stone—ample wood fuel, and good communications by water. At Ching-te-chen the relatively coarse-bodied *shu-fu* ware was developed into a hard white porcelain that no longer reveals the touch of the potter's hand. The practically invisible designs sometimes carved in the translucent body are known as *an-hua,* "secret decoration." In the Yung-lo period (1402–24) the practice began of putting the reign mark on the base. This was first applied to the finest white porcelain and to monochrome ware decorated with copper red under a transparent glaze.

In the early decades of the Ming, the repertoire of designs on Yüan blue-and-white was continued and refined. At first, this ware evidently was considered too vulgar for court use, and none bears the Imperial reign mark until the Hsüan-te period (1425–35). By this time the often crowded Yüan patterns had given way chiefly to dragons or floral motifs of great clarity and grace, vigorously applied in a thick, deep-blue pigment to dishes, vases, stem cups, and flattened pilgrim jars. Sometimes a richer effect was achieved by painting dragons in underglaze red on a blue ground or vice versa. In the Ch'eng-hua period (1464–87), the blue-and-white designs became somewhat tenuous and over-refined, and the characteristic wares made for the Cheng-te Emperor (1505–21) and his Muslim eunuchs often bear Arabic inscriptions. In the Chia-ching (1521–67) and Wan-li (1572–1620) periods, the Imperial kilns



The Cleveland Museum of Art, purchase from the J.H. Wade Fund (59.46)

"River and Mountains on a Clear Autumn Day," hand scroll by Tung Ch'i-ch'ang (1555–1636), Ming dynasty. Ink on paper. In The Cleveland Museum of Art, Ohio, U.S. 37.9 cm × 1.37 m.

were badly mismanaged, and their products were often of poor quality. Private factories, however, turned out lively wares until the end of the dynasty.

"Chicken cups"

Overglaze painting was applied with delicate care in the Ch'eng-hua period, chiefly in the decoration of small wine cups with chicken motifs, much admired by Chinese connoisseurs. Overglaze painting soon became popular, being applied in the 16th century in stronger colours brilliantly contrasted against a dead-white background. These vigorous (*wu ts'ai*) "five-colour" wares were especially free and bold in the Chia-ching and Wan-li periods. Crude but lively imitations of these and of the blue-and-white of Ching-te-chen, made in southern China kilns partly for the Southeast Asian market, are known as "Swatow ware" from one of the export sites. Among the most impressive of Ming pottery types are the *san ts'ai* ("three-colour") wares, chiefly vases and jars decorated with floral motifs in turquoise, purple, yellow, and deep violet blue, the colours being separated by raised lines in imitation of the metal strips used in cloisonné work (see below). This robust ware was made in several centres, the best of it between 1450 and 1550.

Beginning in the early 16th century, a new ceramic tradition emerged in the town of I-hsing, on the western side of Lake T'ai, catering to the tea taste of scholars in the nearby Su-chou area. Individually made, sometimes to order, rather than mass-produced, I-hsing wares were often signed or even poetically inscribed by highly reputable master craftsmen, such as Shih Ta-pin of the Wan-li era and Ch'en Ming-yüan of the Ch'ing dynasty K'ang-hsi period. The wares were usually unglazed and derived their striking colours—brown, beige, reddish purple, yellow, black, and blue—after firing from the distinctive clays of the region and were known as "purple-sand" teapots. Pieces alternated between two body types: complex floral shapes and exquisitely simple geometric designs.

Metalwork. The technique of cloisonné enamel on a metal base, in which the colours are separated by little metal strips, or walls (cloisons), developed in Byzantium about the 5th century AD and probably reached East Asia in the T'ang dynasty. It seems to have been lost again, however, and reintroduced in the 14th century, when it was referred to in the *Ko-ku yao-lun* as "Ta-shih ware" ("Muslim" ware) and "similar to Fo-lang-k'an" (similar to Byzantine enamelware). While some pieces may have been made for the Yüan court, no Chinese Ming cloisonné has been positively identified as of earlier date than the Hsüan-te Emperor's reign (1425–35), when the reign mark first appears. The decoration of the 15th-century pieces is varied, bold, and free; the brass cloisons are thin; and light and dark blue, red, and yellow predominate among the rich and sometimes rather roughly applied colours. Cloisonné of the Ch'ing dynasty and modern times is more perfectly finished than that of the Ming; the shapes are often more ornate but elaborate and monotonous; and the copper cloisons are more insistent in the decoration.

Decoration of 15th-century cloisonné pieces

Lacquerwork. The carved lacquer first developed in the Yüan dynasty continued through the Ming and Ch'ing and was made in many different factories. It reached a high level in carved red lacquer (*t'i-hung*) dishes, trays, covered boxes, and cups of the Yung-lo and Hsüan-te reigns. Yung-lo reign marks, scratched on with a sharp point, are not reliable, but some pieces, bearing carved and gold-inlaid marks of the Hsüan-te Emperor, may be of the period. Decoration of this early Ming lacquer includes both pictorial designs (landscapes with figures in pavilions are common) and rich dragon, phoenix, and floral motifs, carved deeply in a full, freely flowing and plastic style, often against a yellow background. While this style continued into the 16th century, the Chia-ching period also saw the emergence of more realistic and intricate designs that are shallower and more sharply carved, sometimes through as many as nine layers of different colours, on a background consisting of minute brocade (allover floral and figure designs) or diaper (diamond-shaped) patterns.

Ch'ing dynasty (1644–1911/12). The Manchu conquest did not produce a dislocation of Chinese social and cultural life as had the Mongol invasion. On the contrary, even before their conquest, the Manchus began imitating Chinese ways, and the Ch'ing rulers, particularly K'ang-hsi (1661–1722) and Ch'ien-lung (1735–96), were well-educated men, eager to enlist the support of Chinese scholars. They were extremely conservative in their political and cultural attitudes; and in artistic taste, their native love of extravagance (which the Chinese viewed as "barbarous") was tempered, ironically, by an equally strong conservative propensity. The art of the Ch'ing dynasty, even the painting of many of its finest eccentrics and the design of its best gardens, is similarly characterized both by lavish decoration and ornate effects as well as by superb technique and conservative taste. In the 19th century, however, China's internal weakness and humiliation by the Western powers were reflected in a growing stagnation of the arts.

Architecture. Ch'ing dynasty work in the Forbidden City was confined chiefly to the restoration or reconstruction of major Ming buildings, although the results were typically more ornate in detail and garish in colour than ever before. The Manchu rulers were most lavish in their summer palaces, created to escape the heat of the city. In 1703 the K'ang-hsi Emperor began near the old Manchu capital, Cheng-te, a series of palaces and pavilions set in a natural landscape. Engravings of these made by the Jesuit father Matteo Ripa in 1712–13 and brought by him to London in 1724 are thought to have influenced the revolution in garden design that began in Europe at about this time. Near the Cheng-te palace were built several imposing Buddhist temples in a mixed Sino-Tibetan style that reflects the Tibetan Buddhist leanings of the K'ang-hsi, Yung-cheng, and Ch'ien-lung emperors.

Summer palaces

About 1687 the K'ang-hsi Emperor had begun to create another garden park northwest of Peking, which grew under his successors into the enormous Yüan-ming Yüan ("Garden of Pure Light"). Here were scattered a great number of official and palace buildings, to which the Ch'ien-lung Emperor moved his court semipermanently. In the northern corners of the Yüan-ming Yüan, the Jesuit missionary and artist Giuseppe Castiglione (known in China as Lang Shih-ning) designed for Ch'ien-lung a series of extraordinary Sino-Rococo buildings, set in Italianate gardens ornamented with mechanical fountains designed by the Jesuit priest Michel Benoist. Today the Yüan-ming Yüan has almost completely disappeared, the foreign-style buildings having been burned by the French and British in 1860. To replace it, the empress dowager Tz'u-hsi greatly enlarged the new summer palace (I-ho Yüan) along the shore of K'un-ming Lake to the north of the city.

The finest architectural achievement of the period, however, occurred in private rather than institutional architecture—namely, in the scholars' gardens of southeastern China, in such towns as Su-chou, Yang-chou, and Wu-hsi. As these often involved renovations carried out on Yüan and Ming dynasty foundations, it remains difficult to discern the precise outlines of their innovations. With the aid of paintings and Chi Ch'eng's *Yüan yeh* ("Forging a Garden," 1631–34), it becomes evident that, as in the worst of Ch'ing architecture, these gardens became ever more ornate. The best examples, however, remain well within the bounds of good taste owing to the cultivated sensibility of scholarly taste, and they were distinguished by an inventive imagination lacking in Manchu court architecture. Such gardens were primarily Taoist in nature, intended as microcosms invested with the capacity to engender tranquillity and induce longevity in those who lodged there. The chief hallmark of these gardens was the combination of a central pond, encompassing all the virtues of yin in the Chinese philosophical system, with the extensive use of rugged and convoluted rockery, yang, which summed up the lasting Chinese adoration of great mountain chains through which flowed the vital energy of the earth. (The most precious rocks were harvested from the bottom of Lake T'ai near Su-chou.)

Private scholars' gardens

Throughout this urban garden tradition, where the scale was necessarily small and space was strictly confined, designers attempted to convey the sense of nature's vastness by breaking the limited space available into still smaller but ever varied units, suggesting what could not otherwise literally be obtained, and expanding the viewer's imagina-

tion through inventive architectural design. Among those gardens still preserved today, the Liu Garden in Su-chou offers the finest general design and the best examples of garden rockery and latticed windows, while the small and delicate Garden of the Master of Nets (Wang-shih Yüan), also in Su-chou, provides knowledgeable viewers a remarkable series of sophisticated surprises.

Painting and printmaking. The dual attraction of the Manchu rulers to unbridled decoration and to orthodox academicism characterized their patronage at court. In regard to the former, they favoured artists such as Yüan Chiang, who, in the reign of K'ang-hsi, applied to the model of Kuo Hsi, with great decorative skill, the mannered distortions that had cropped up in the late Ming partly as a result of Ming artists' exposure to an unfamiliar Western art. More thoroughly Westernized work, highly exotic from the Asian perspective, was produced both by native court artists like Chiao Ping-chen, who applied Western perspective to his illustrations of *Keng-chih-t'u* ("Rice and Silk Culture"), which were reproduced and distributed in the form of wood engravings in 1696, and by the Jesuit missionary and painter Giuseppe Castiglione. In the mid-18th century Castiglione produced a Sino-European technique that had considerable influence on court artists such as Tsou I-kuei, but he was ignored by literati critics. His depictions of Manchu hunts and battles provide a valuable visual record of the times.

Ch'ing patronage of the scholar-amateur style of painting

On the other hand, Manchu emperors saw to it that conservative works in the scholar-amateur style by Wang Hui, Wang Yüan-ch'i, and other followers of Tung Ch'i-ch'ang were also well represented at court, largely putting an end to the conflict at court between professional and amateur styles that had been introduced in the Sung and played a significant role in the Ming. In a sense, the amateur style was crowned victor, but it came at the expense of the amateurism that had defined its purpose. This politically effective aspect of Manchu patronage may have been less a specifically calculated strategy than a natural extension of their concerted attempts to cultivate and recruit the scholar class in order to establish their legitimacy.

The Ch'ien-lung Emperor was the most energetic of royal art patrons since Hui-tsung of the Sung, building an Imperial collection of over 4,000 pre-Ch'ing paintings and calligraphy and cataloging them in successive editions of the *Shih-ch'ü pao-chi.* The shortcomings of his taste, however, were displayed in his preference for recent forgeries rather than the originals in his collection (notably, copies of Huang Kung-wang's "Dwelling in the Fu-ch'un Mountains" and of Fan K'uan's "Travelers Among Mountains and Streams") and in his propensity for covering his collected masterpieces with multiple impressions of court seals and calligraphic inscriptions in a mediocre hand.

Sen-oku Hakko kan, Kyoto



"Two Birds," album leaf by Chu Ta (Pa-ta Shan-jen), 1694, Ch'ing dynasty. Ink and slight colour on paper. In the Sumitomo Collection, Sen-oku Hakko kan, Kyōto, Japan. 31.7 × 27.7 cm.

The conservatism of Ch'ing period painting was exemplified by the so-called "Six Masters" of the late 17th–early 18th century, including the so-called "Four Wangs," Wu Li, and Yün Shou-p'ing. In the works of most of these artists and of those who followed their lead, composition became routinized, with little in the way of variation or genre detail to appeal to the imagination; fluency of execution in brushwork became the exclusive basis for appreciation. Wang Shih-min, who had been a pupil of Tung Ch'i-ch'ang, retired to T'ai-ts'ang near modern Shanghai at the fall of the Ming, making it the centre of a school of scholarly landscape painting that included his friend Wang Chien and the younger artist Wang Hui. Wang Hui was a dazzling prodigy whose landscapes included successful forgeries of Northern Sung and Yüan masters and who did not hesitate to market the "amateur" practice, both among fellow scholars and at the Manchu court; however, the hardening of his style in his later years foreshadowed the decline of Ch'ing literati painting for lack of flexible innovation. By contrast, Wang Shih-min's grandson, Wang Yüan-ch'i, was the only one of these six orthodox masters who fully lived up to Tung Ch'i-ch'ang's injunction to transform the styles of past models creatively. At court, Wang Yüan-ch'i rose to high office under the K'ang-hsi Emperor and served as chief compiler of the Imperial painting and calligraphy catalog, the *P'ei-wen-chai shu-hua-p'u.*

The "Individualists"

Receiving no patronage from the Manchu court and leaving only a minor following before the latter half of the 19th century was a different group of artists, now frequently referred to as "Individualists." Collectively, these artists represent a triumphant, if short-lived, moment in the history of literati painting, triggered in good part by the emotionally cathartic conquest of China by the Manchus. They shared in common a rejection of Manchu political authority and the choice of an eremitic, often impoverished lifestyle that obliged them to trade their works for their sustenance in spite of their allegiance to amateur ideals. Stylistically, just like their more orthodox contemporaries, they often revealed the influence of Tung Ch'i-ch'ang's systematization of painting method; but, unlike the more conservative masters, they pursued an emotional appeal reflective of their own temperaments. For example, Kung Hsien, a Nanking artist whose budding political career was cut short by the Manchu conquest, used repetitive forms and strong tonal contrasts to convey a pervasive feeling of repressive constraint, lonely isolation, and gloom in his landscapes. He was the most prominent of the artists who came to be known as the "Eight Masters of Nanking." This group was only loosely related stylistically, though contemporary painters from Nanking did share solidity of form derived from Sung prototypes and, possibly, from the influence of Western art.

The landscapes of K'un-ts'an (or Shih-ch'i), who became a somewhat misanthropic abbot at a Buddhist monastery near Nanking, also express a feeling of melancholy. His works were typically inspired by the densely tangled brushwork of Wang Meng of the Yüan.

Another Individualist artist to join the Buddhist ranks was Hung-jen, exemplar of a style that arose in the Hsin-an or Huichou district of southeastern Anhwei province and drew on the famed landscape of the nearby Huang Mountains. The group of artists now known as the "Anhwei School" (including Ting Yün-p'eng, Hsiao Yün-ts'ung, Mei Ch'ing, Cha Shih-piao, and Tai Pen-hsiao) mostly pursued an opposite emotional extreme from Kung Hsien and K'un-ts'an, a severe coolness based on the sparse, dry linear style of the Yüan artist Ni Tsan. However individualistic, virtually all these artists reveal the influence of Tung Ch'i-ch'ang's compositional means.

Two artists, both members of the deposed and decimated Ming royal family, stood out among these Individualist masters and left, albeit belatedly recognized, the most enduring legacy of all. Known by a sequence of names,

The painters Chu Ta and Tao-chi

perhaps designed to protect his royal identity, Chu Ta, or Pa-ta Shan-jen, suffered or at least feigned a period of madness and muteness in the 1680s. He emerged from this with an eccentric style remarkable for its facility with extremes, alternating between a wet and wild manner and a dry, withdrawn use of brush and ink. His paintings of glowering birds and fish casting strange and ironic glances, as well as his structurally interwoven studies of rocks and vegetation, are virtually without precedent in composition, although aspects of both the eccentric Hsü Wei and Tung Ch'i-ch'ang are discernible in his work. His esoteric inscriptions reveal a controlled intent rather than sheer lunacy and suggest a knowledgeable, if hard to unravel, commentary on China's contemporary predicament.

Chu Ta's cousin Tao-chi was raised in secret in a Ch'an Buddhist community. He traveled widely as an adult in such varied artistic regions as the Huang Mountains district of Anhwei province and Nanking and finally settled in the newly prosperous city of Yang-chou, where in his later years he publicly acknowledged his royal identity, renounced his Buddhist status, and engaged in professional practices. His work has a freshness inspired not by masters of the past but by an unfettered imagination, with brush techniques that were free and unconventional and a daring use of colour. In his essay *Hua yü lu* ("Comments on Painting"), he ridiculed traditionalism, writing that his own method was "no method" and insisting that, like nature, creativity with the brush must be spontaneous and seamless, based on the concept of *i-hua*, the "unifying line."

Tao-chi's extreme stand in favour of artistic individuality stands out against the growing scholasticism of Ch'ing painting and was an inspiration to the artists, roughly grouped together as the "Eight Eccentrics" (including Cheng Hsieh, Hua Yen, Huang Shen, Kao Feng-han, Chin Nung, and Lo P'ing), who were patronized by the rich merchants in early 18th-century Yang-chou. The art of Chu Ta and Tao-chi was not firmly enshrined, however, until the late 19th century, when a new individualist thrust appeared in Shanghai in response to the challenge of Western culture. Their influence on Chinese art since then, especially in the 20th century, has been profound.

Calligraphy. Those who straddled the Ming-Ch'ing transition continued the exploration of individualistic calligraphic styles (particularly in cursive and semicursive scripts) and included many who also excelled at painting, such as Fu Shan, Fa Jo-chen, Chu Ta, and Tao-chi. The chief contribution of Ch'ing calligraphers came later, however, with a resurgence in the importance of seal and clerical scripts (which had survived primarily in seal carving and in the writing of titled frontispieces for painted hand scrolls). This was based on a renewed scholarly interest in inscribed Chou dynasty bronzes and in northern stelae of the Han, Six Dynasties, and T'ang periods. While an indication of Chinese calligraphers' interest in seal and clerical type scripts can already be seen in the early Ch'ing writings of Fu Shan, Chu Ta, and Tao-chi, a more scientific study of bronze and stone inscriptions (*chin-shih hsüeh*) was begun in the early to mid-18th century by such scholar-calligraphers as Wang Shu. Thereafter, practitioners could be divided among imitative calligraphers and those who were more creative in their adaptation of these ancient scripts. Among the latter were the 18th-century Yang-chou painters Chin Nung and Kao Feng-han, as well as the 18th- and 19th-century calligraphers Teng Shih-ju, I Ping-shou, Juan Yüan, and Wu Ta-ch'eng. The interest in ancient styles continued in the early 20th century, when the paleographer Tung Tso-pin and other calligraphers pioneered the adaptation to the modern brush of the script used to inscribe Shang dynasty divinations, which had recently been excavated.

Ceramics. The pottery industry suffered severely in the chaotic middle decades of the 17th century, of which the typical products are "transitional wares," chiefly blue-and-white. The Imperial kilns at Ching-te-chen were destroyed and were not fully reestablished until 1682, when the K'ang-hsi Emperor appointed Ts'ang Ying-hsüan as director. Under his control, Imperial porcelain reached a level of excellence it had not seen for well over a century.

Imperial porcelain



"Hut at the Foot of Mountains," album leaf by Tao-chi, *c.* 1695, Ch'ing dynasty. Ink and colour on paper. In the C.C. Wang Family Collection, New York City. 24.1 × 27.9 cm.
From the C.C. Wang Family Collection, New York City

The finest pieces include small monochromes, which recaptured the perfection of form and glaze of classic Sung wares. New colours and glaze effects were introduced, such as eel-skin yellow, snakeskin green, turquoise blue, and an exquisite soft red glaze shading to green (known as "peach-bloom") that was used for small vessels made for the scholar's desk. Also perfected was *lang-yao* (sang-de-boeuf, or oxblood, ware), which was covered with a rich copper-red glaze. K'ang-hsi period blue-and-white is particularly noted for a new precision in the drawing and the use of cobalt washes of vivid intensity.

Five-colour (*wu ts'ai*) overglaze painted wares of the K'ang-hsi period became known in Europe as famille verte from the predominant green colour in their floral decoration. These wares also included expert imitations of overglaze painting of the Ch'eng-hua Emperor's reign. Another variety has floral decoration painted directly on the biscuit (unglazed pottery body) against a rich black background (famille noire). Toward the end of the K'ang-hsi reign, a rose pink made from gold chloride was introduced from Europe. It was used with other colours in the decoration of porcelain (famille rose) and in cloisonné and overglaze painting.

Famille rose porcelain reached a climax of perfection at Ching-te-chen under the direction of Nien Hsi-yao (1726–36) and continued with scarcely diminishing delicacy through the Ch'ien-lung period. Meanwhile, the skill of the Ching-te-chen potters was being increasingly challenged by the demand at court for imitations in porcelain of archaic bronzes, gold, and jade and for such inappropriate objects as musical instruments and perforated and revolving boxes, which were highly unsuited to manufacture in porcelain. Although fine porcelain was made from time to time in the 19th century, notably in the Tao-kuang and Kuang-hsü reigns, the quality as a whole greatly declined.

Textiles. Ming and Ch'ing textiles fully display the Chinese love of pageantry, colour, and fine craftsmanship. Prominent among woven textile patterns are flowers and dragons against a background of geometric motifs that date to the late Chou and Han. Ch'ing robes were basically of three types. The *ch'ao-fu* was a very elaborate court ceremonial dress; the emperor's robe was adorned with the auspicious Twelve Symbols described in ancient ritual texts, while princes and high officials were allowed nine symbols or fewer according to rank. The *ts'ai-fu* ("coloured dress"), or "dragon robe," was a semiformal court dress in which the dominant element was the Imperial five-clawed dragon (*lung*) or the four-clawed dragon (*mang*). In spite of repeated sumptuary laws issued during the Ming and Ch'ing, the five-clawed dragon was seldom reserved for

Three types of Ch'ing robe

objects of exclusively Imperial use. Symbols used on the dragon robes also included the eight Buddhist symbols, symbols of the Taoist Eight Immortals, eight precious things, and other auspicious devices. "Mandarin squares" had been attached front and back to Ming official robes as symbols of civil and military rank and were adapted by the Manchus to their own distinctive dress.

Jade and small-scale carving. China directly controlled the Central Asian jade-yielding regions of Ho-tien and Yarkand between about 1760 and 1820, during which time much fine nephrite was sent to Peking for carving. Jadeite from Myanmar (Burma) reached the capital from the second quarter of the 18th century, and chromite- or graphite-flecked "spinach jade" from the Baikal region of Siberia was imported in the 19th century. The finest Ch'ing dynasty jade carving is often assigned to the reign of Ch'ien-lung, but carved jade is difficult to date. Typical of what is considered of Ch'ien-lung date are vases with lids and chains carved from a single block, vessels in antique bronze shapes with pseudo-archaic decoration, fairy mountains, and brush pots for the scholar's desk.

The same forms and motifs were also skillfully employed in the carving of other ornamental minerals, such as rock crystal, rose quartz, agate, lapis lazuli, and soapstone. Owing to the early perfecting of porcelain in China, glassmaking never became a major craft. In the 18th century, however, it was somewhat developed, partly under Western influence, the chief centre of production for the court being Po-shan in Shantung. Other minor crafts that achieved a high level of technical virtuosity in the Ch'ing dynasty include the carving of birch root, bamboo root, rhinoceros horn, elephant ivory, and hornbill ivory. Snuff bottles were made from a variety of hard stones, porcelain, glass, coral, cloisonné, and other materials. The technique of painting on the inside of glass snuff bottles was developed in the second half of the 19th century.

Other minor crafts

Modern period (since 1912). The arts of China since 1912 have reflected the emergence of China into the modern world, the impact of Western art, and the political and military struggles of the period, culminating in the war with Japan (1937–45) and the civil war that ended in the establishment in 1949 of the People's Republic of China.

Architecture. Until the mid-1920s official and commercial architecture were chiefly in the eclectic European style of such treaty ports as Canton, Amoy, Fu-chou, and Shanghai, much of it designed by foreign architects. In 1925, however, a group of foreign-trained Chinese architects launched a renaissance movement to study and revive traditional Chinese architecture and to find ways of adapting it to modern needs and techniques. In 1930 they founded Chung-kuo Ying-tsao Hsüeh-she, the Society for the Study of Chinese Architecture, which was joined in the following year by Liang Ssu-ch'eng, the dominant figure in the movement for the next 30 years. The fruits of their work can be seen in new universities and in major government and municipal buildings in Nanking and Shanghai. The war with Japan put an end to further developments along these lines, however. Since 1949, Peking and other big cities have been transformed by spectacular planning projects, but an awareness of the traditional role of symbolism in architecture has been retained and adapted to communist propaganda purposes. Large portions of the Forbidden City in Peking have been restored and established as a public museum, but a section has been given over to residences for the new ruling elite. A new primary thoroughfare (Ch'ang-an Boulevard) has been established, running east and west in front of the old palaces, contrary to the old north-south axis. A vast square for public political activity has been created in front of the Gate of Heavenly Peace (T'ien-an Men, entryway to the Imperial City), flanked on one side by the Museum of Chinese History and the Revolution and on the other by the Great Hall of the People, built in Soviet style in 1959 during the Great Leap Forward. Most of the city's magnificent walls were torn down before or during the Cultural Revolution on the pretext that they impeded the flow of traffic. Finally, the regime's founder, Mao Zedong, who died in 1976, was buried in a mausoleum bearing a striking resemblance to the Lincoln Memorial in Washington, D.C.; the tomb is located in the centre of the city at the south end of Tien-anmen Square, where it obstructs the north-south axis in flagrant violation of traditional geomantic principles.

Urban planning projects

Painting and printmaking. Shanghai, which had been forcibly opened to the West in 1842 and boasted a newly wealthy clientele, was the logical site for the first modernist innovations. A Shanghai regional style appeared by the 1850s, led by Jen Hsiung, his more popular follower Jen I (or Jen Po-nien), and Jen I's follower Wu Ch'ang-shih. It drew its inspiration from a series of Individualist artists of the Ming and Ch'ing, including Hsü Wei, Ch'en Shun, Ch'en Hung-shou, Chu Ta, and Tao-chi; it focused on birds and flowers and figural themes more than the old landscape tradition; and it emphasized decorative qualities, exaggerated stylization, and satiric humour rather than refined brushwork and sober classicism. Under Wu Ch'ang-shih's influence, this style was passed on to Peking through the art of Ch'en Heng-k'o (or Ch'en Shih-tseng) and Ch'i Pai-shih in the first years after the revolution.

The Japanese faced the issues of modernization earlier than the Chinese, blending native and Western traditions in Nihonga painting and establishing an institutional basis of support under the leadership of Okakura Kakuzō, who founded the Tokyo Fine Arts School in 1889. Thus, it is not surprising that among the first in China to respond similarly were artists who had traveled to Japan, including Kao Chien-fu, his brother Kao Ch'i-feng, and Ch'en Shu-jen. Inspired by the "New Japanese Style," the Kao brothers and Ch'en inaugurated a "New National Painting" movement, which in turn gave rise to a Cantonese (or "Ling-nan") regional style that incorporated Euro-Japanese characteristics. Although the new style did not produce satisfying or lasting solutions, it was a significant harbinger and continued to thrive in Hong Kong, practiced by such artists as Chao Shao-ang. The first establishment of Western-style art instruction also dates from this period. A small art department was opened in Nanking High Normal School in 1906, and the first art academy, later to become the Shanghai Art School, was founded in the year of the revolution, 1911, by the 16-year-old Liu Hai-su, who in the next decade pioneered the first public exhibitions (1913) and the use of live models, first clothed and then nude, in the classroom.

Establishment of Western-style art schools

Increasingly, by the mid-1920s, young Chinese artists were attracted not just to Japan but also to Paris and German art centres. A trio of these artists brought back some understanding of the essential contemporary European traditions and movements. Liu Hai-su was first attracted to Impressionist art, while Lin Feng-mien, who became director of the National Academy of Art in Hang-chou in 1928, was inspired by Postimpressionist experiments in colour and pattern by Henri Matisse and the Fauvists. Lin advocated a synthesis combining Western techniques and Chinese expressiveness and left a lasting mark on the modern Chinese use of the brush. Hsü Pei-hung, head of the National Central University art department in Nanking, eschewed European modernist movements in favour of Parisian academism. He developed his facility in drawing and oils, later learning to imitate pencil and chalk with the Chinese brush; the monumental figure paintings he created served as a basis for Socialist Realist painters after the communist revolution of 1949. By the 1930s, all these modern trends were clearly developed and institutionalized. Although most of the major artists of the time advocated modernism, two continued to support more traditional styles: Ch'i Pai-shih, who combined Shanghai style with an infusion of folk-derived vitality, and the relatively conservative landscapist Huang Pin-hung, who demonstrated that the old tradition could still produce a rival to the great masters of the 17th century.

Socialism produced a new set of artistic demands that were first met not by painting but by the inexpensive mass medium of woodblock prints (which had been invented in China and first used in the T'ang dynasty to illustrate Buddhist sutras). Initially stimulated by the satiric leftist writer Lu Hsün, printmakers flourished during the 1930s and '40s under the dual influence of European socialist artists like Käthe Kollwitz and the Chinese folk tradition of New Year's prints and papercuts. Among the most

prominent print artists were Li Hua and Ku Yüan, who attained a new standard of political realism in Chinese art.

Aesthetic principles of the Communist Party

In 1942, as part of the Chinese Communist Party's first intellectual rectification movement, Mao Zedong delivered two speeches at the Yen-an Forum on Literature and Art that became the official party dictates on aesthetics for decades to come. Mao emphasized the subordination of art to political ends, the necessity for popularization of styles and subjects in order to reach a mass audience, the need for artists to share in the lives of ordinary people, and the requirement that the party and its goals be treated positively rather than subjected to satiric criticism. "Art for art's sake" was strictly denounced as a bourgeois liberal attempt to escape from the truly political nature of art. Although Mao later defended a place for the artistic study of nude models, a staple of Western naturalism, the tone he set led to severe limitations in the actual practice of this.

The Sino-Japanese War of 1937–45 led many artists of varied persuasions to flee eastern China for the temporary Nationalist capital in Chungking, Szechwan province, bringing a tremendous mixing of styles and artistic ferment, but the opportunity for innovation which this promised was thwarted by subsequent events. After the 1949 revolution, Communist Party control of the arts was firmly established by the placement of the academies under the jurisdiction of the Ministry of Culture, by the creation of artists' federations and associations (which served as an exclusive pathway to participation in exhibitions and other means of advancement) under the management of the party's Department of Propaganda, by the establishment of a strict system of control over publications, and by the virtual elimination of the commercial market for contemporary arts.

Throughout the 1950s, as Socialist Realist standards were gradually implemented, oil painting and woodblock printing were favoured and political cartoons and posters were raised to artistic status. Artists in the traditional media—with their basis in the Individualist art of the old "feudal" aristocracy—struggled institutionally for survival, eventually succeeding only as a result of the nationalist fervour that accompanied China's ideological break with the Soviet Union late in the decade. The internationalist but relatively conservative Hsü Pei-hung was installed as head of the new Central Academy of Fine Arts in Peking but died in 1953. Other older-generation leaders passed on shortly afterward (Ch'i Pai-shih and Huang Pin-hung) or were shunted aside (Liu Hai-su and Lin Feng-mien), and a younger generation soon came to the fore, ready to make the necessary compromises with the new regime. The talented landscapist Li Keran, who had studied with Ch'i Pai-shih, Lin Feng-mien, Huang Pin-hung, and Hsü Pei-hung, combined their influences with realistic sketching to achieve a new naturalism in the traditional medium. A leading figure painter was Cheng Shifa, a descendant of the Shanghai school who brought that style to bear in politically polished depictions of China's minority peoples. Many talented artists, including Luo Gongliu and Ai Zhongxin, painted in oils, which, because of their link to the Soviet Union and Soviet art advisors, held a favoured position until the Sino-Soviet split of the late 1950s.

The Cultural Revolution

While the early 1960s provided a moment of political relaxation for Chinese artists, the Cultural Revolution of 1966–76 brought unprecedented hardships, ranging from forced labour and severe confinement to death. Destruction of traditional arts was especially rampant in the early years of the movement. Only those arts approved by a military-run apparatus under the sway of Mao's wife, Jiang Qing, could thrive; these followed the party's increasingly strict propagandist dictates and were often created anonymously as collective works.

The passing of Mao and Maoism after 1976 brought a new and sometimes refreshing chapter in the arts under the leadership of Deng Xiaoping. The 1980s were characterized by decreasing government control of the arts and increasingly bold artistic experimentation. Three phenomena in 1979 announced this new era: the appearance of Cubist and other Western styles, as well as nude figures, in the murals publicly commissioned for the new Peking airport (although the government "temporarily" covered the nudes); a private arts exhibition by The Stars art group at the Peking Art Gallery; and the rise of a truly realistic oil painting movement, which swept away the artificiality of Socialist Realist propaganda. A resurgence of traditional Chinese painting occurred in the 1980s, featuring the return of formerly disgraced artists, including Li Keran, Cheng Shifa, Shi Lu, and Huang Yongyu, and the emergence of such fresh talents as Wu Guanzhong, Jia Youfu in Sian, and Li Huasheng of Szechwan province.



"Fleeing Refugees," woodblock print by Li Hua, 1944. Ink on paper. 21 × 4.6 cm.
© Li Hua/ChinaStock Photo Library

After 1985, as an increasingly bold avant-garde movement arose, the once-threatening traditional-style painting came to seem to the government like a safe alternative. In the final months before the June 1989 imposition of martial law in Peking, an exhibition of nude oil paintings from the Central Academy of Fine Arts at the Chinese National Gallery and an avant-garde exhibition featuring installation art, performance art (which lacked the necessary permit for a public gathering), and the mockery of the government through printed scrolls full of unreadable pseudo-characters drew record crowds. The latter was closed by police, and both exhibits were eventually denounced as having lowered local morals, helping to precipitate the tragic events that followed in June 1989. New limitations on artistic production, exhibition, and publication ensued. At the conclusion of these events, a number of leading artists, including Huang Yongyu, fled China, joining others who had previously fled or abandoned China to establish centres of Chinese art throughout the world. Among the leading Chinese artists outside mainland China have been Chao Wu-chi (Zao Wou-ki), who settled in Paris; Chang Dai-chien in Taiwan, Brazil, and the United States; Wang Chi-ch'ien (C.C. Wang) and Tseng Yu-ho in the United States; Liu Kuo-sung, Ch'en Ch'i-kwan, and Ho Huai-shuo in Taiwan; Fang Chao-lin in Hong Kong and London; and Lin Feng-mien in Hong Kong. (M.Su./Je.Si.)

Korean visual arts

GENERAL CHARACTERISTICS

The art produced by peoples living in the peninsula of Korea has traditionally shared aesthetic concepts, motifs, techniques, and forms with the art of China and Japan. Yet it has developed a distinctive style of its own. In general, Korean art has neither the grandeur and aloofness of Chinese art nor the decorative sophistication of Japanese art, and Korean artists were often not as technically perfect or precise as their Chinese and Japanese counterparts. Instead, the beauty of Korean art and the strength of its artists lay in simplicity, spontaneity, and a feeling of harmony with nature. In mood, the art of Korea is often characterized by a sense of loneliness arising from the serenity of the image and reflecting the Korean philosophy of resignation.

The basic trend of Korean art has been naturalistic, a

characteristic already evident as early as the Three Kingdoms period (*c.* 57 BC–AD 668) but fully established by the Unified, or Great, Silla (Korean: Shinla) period (668–935). The traditional attitude of accepting nature as it is resulted in a highly developed appreciation for the simple and unadorned. Korean artists, for example, favour the unadorned beauty of raw materials, such as the natural patterns of wood grains. The Korean potter was characteristically unconcerned about mechanical perfection of his surfaces, curves, or shapes. His concern was to bring out the inherent or natural characteristics of his materials and the medium. Potters, therefore, were able to work unself-consciously and naturally, producing wares of engaging simplicity and artistic distinctiveness.

Simplicity applied not only to economy of shape but also to the use of decorative motifs and devices. The intervention of the human hand is restricted to a minimum in Korean art. A single stem of a flower, for instance, may be drawn in a subtle shade of blue on the side of a white porcelain vase or bottle, but never merely from a desire to fill an empty space. The effect is rather to enlarge the white background.

The avoidance of extremes is another characteristic tradition in Korean art. Extreme straightness of line was disliked as much as extreme curvilinearism. The straight bold contour of a Sung dynasty (960–1279) Chinese bowl becomes a graceful, modest curve in a Korean bowl of the Koryŏ period (918–1392). The sharply curving Chinese roof is modified into a gently sloping roof. Sharp angles, strong lines, steep planes, and garish colours are all avoided. The overall effect of a piece of Korean art is generally gentle and mellow. It is an art of fluent lines. What is most striking is not the rhythm so much as the quiet inner harmony.

STYLISTIC AND HISTORICAL DEVELOPMENT

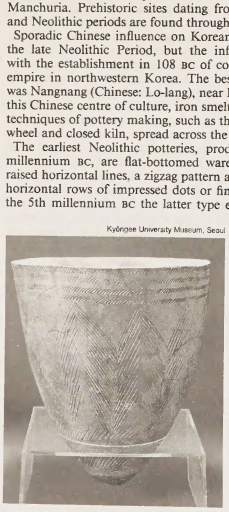
The formative period. Both archaeological and linguistic evidence indicates that the Korean people originally spread into the Korean peninsula from Siberia by way of Manchuria. Prehistoric sites dating from the Paleolithic and Neolithic periods are found throughout the peninsula.

Sporadic Chinese influence on Korean culture began in the late Neolithic Period, but the influence intensified with the establishment in 108 BC of colonies of the Han empire in northwestern Korea. The best known of these was Nangnang (Chinese: Lo-lang), near P'yŏngyang. From this Chinese centre of culture, iron smelting and advanced techniques of pottery making, such as the use of a potter's wheel and closed kiln, spread across the peninsula.

Comb-pattern pottery

The earliest Neolithic potteries, produced in the 6th millennium BC, are flat-bottomed wares decorated with raised horizontal lines, a zigzag pattern around the rim, or horizontal rows of impressed dots or fingernail marks. In the 5th millennium BC the latter type evolved into what is known as comb-pattern pottery, which characteristically features a pointed or rounded bottom and overall geometric patterns of herringbone, meander, and concentric semicircles, produced by incised, impressed, or dragged dots and short lines. The linear, abstract tendency of these Neolithic potteries basically falls in the tradition of prehistoric Siberian art.

Kyŏngee University Museum, Seoul



Neolithic comb-pattern pottery, from Amsadong, Seoul, *c.* 4th millenium BC. In the Kyŏnghee University Museum, Seoul. Height 40.5 cm.

In the ensuing Bronze Age (*c.* 1000–300 BC) and Early Iron Age (*c.* 300–0 BC) more types of pottery of improved quality appeared. Painted pieces derived from Chinese painted pottery have been found in North Korea, while wares devoid of surface decoration were used in other areas of the peninsula. Clay, bone, or stone figurines of seated or standing shamanistic deities were also produced.

It was also during this time that bronze- and iron-working centres were established in Korea. Bronze daggers, mirrors, and perforated pole finials, all ultimately of Siberian origin, were cast. The daggers are of the type widely used by the Scythian peoples of the Eurasian steppe. The mirrors were also of a non-Chinese type, with twin knobs placed a little off centre against a tightly composed, geometric design made up of finely hatched triangles.

More evidence of the Siberian art tradition in prehistoric Korea can be seen in a rock-cut drawing discovered in 1970 at Pan'gudae, near the southeastern coast of South Korea. Pecked line drawings and silhouettes of animals, including whales, dolphins, tigers, wolves, and deer, are depicted on a large (8 by 2 metres), smooth vertical surface of the rock. Some of the animals have a "life line" drawn from the mouth to the anus in the so-called X-ray style of Siberian rock art.

Three Kingdoms period (*c.* 57 BC–AD 668). The first major period of Korean art during recorded history is the period of the Three Kingdoms (*c.* 57 BC–AD 668), when the peninsula of Korea was ruled by three monarchies. The Koguryŏ kingdom (traditionally dated 37 BC–AD 668) was the northernmost of the three, both geographically and culturally. First established in southern Manchuria, its lifestyle was based on the typically austere cultural patterns of northern Asia, evolved in a region characterized by its scarcity of arable land and severity of climate. The Paekche kingdom (traditionally dated 18 BC–AD 660) was centred in southwestern Korea, south of the present-day city of Seoul. This was a favourable geographic position for receiving foreign cultural influences. Paekche art, therefore, was open and receptive to Chinese influences. Northern Chinese cultural elements were introduced by land through the Koguryŏ kingdom, while southern Chinese influences easily crossed the navigable East Asian seas. The kingdom of Silla (traditionally dated 57 BC–AD 668) was the oldest of the monarchies. It originated in the present city of Kyŏngju and eventually came to cover most of southeastern Korea east of the Naktong River. The original territory of the Silla kingdom, the modern North Kyŏngsang province, is a mountain-secluded triangle, a geographic factor sometimes offered as an explanation for the distinctiveness and conservatism of its art.

The kingdoms of Koguryŏ, Paekche, and Silla

The introduction of Buddhism into Koguryŏ from China (AD 372) brought a sudden efflorescence of the arts. Until that time a sustained tradition of large-scale art had been virtually nonexistent. The Koguryŏ kings started the building of temples and pagodas, and sculpture, in the form of Buddha images, made its appearance. By the 6th century, the Silla and Paekche kings had also become converts to the new faith, and from then on Buddhism remained the main inspiration of Korean art until the 15th century.

During the Three Kingdoms period there were three political and cultural centres: P'yŏngyang, the capital of Koguryŏ, in the northwest; the Kongju-Puyŏ region, the Paekche heartland, in the southwest; and Kyŏngju, the capital of Silla, in the southeast. Silla and Paekche, along with the minor state of Kaya (also known as Kara or Karak; Japanese: Mimana) in the south, maintained close cultural contacts with Japan, and it was at this time that the significant Korean influence on Japanese art began. The Paekche kingdom first introduced Buddhism and Chinese writing to Japan. South Korean immigrants to Japan founded important centres of learning and the arts. The Sue pottery of the Tumulus, or Kofun, period (also

known as the Great Burial Period) was the Japanese version of the Silla pottery of Korea. Even the famed wall paintings of the Hōryū Temple in Nara, Japan, have been attributed to a northern Korean painter, Tamjing, from the Koguryŏ kingdom.

Except for several small Buddhist images in bronze and clay, very little remains of Koguryŏ's religious art. A considerable amount, however, has been preserved from the two southern kingdoms. Paekche was the first to use granite in the construction of pagodas and sculpture. After the Three Kingdoms period, granite, which is abundant in Korea, was widely used in construction and sculpture. The granite pagodas of Korea stand in sharp contrast to the brick pagodas of China and the wooden pagodas of Japan.

Earliest uses of granite

The surviving secular art of the period consists chiefly of burial gifts taken from tombs. Not much is available from Koguryŏ and Paekche, because the tombs were too easily accessible and have long since been looted. However, much pottery, along with items used for personal adornment, has continued to turn up in the second half of the 20th century from the less accessible Silla tombs. The most valuable pieces of Old Silla art came from huge mounded tombs in the Kyŏngju area. The rich Silla gold mines, exhaustively worked, yielded the abundance of gold ornaments reflected in the ancient Japanese epithet Manokagayaku Shiragi, or "Eye-Brightening Silla."

Architecture. No original examples of Koguryŏ architecture remain, except for some foundation stones that vaguely suggest a building site, possibly of a royal villa, on the Yalu River near Chi-an, China. In the P'yŏngyang area, three temple sites, probably of the late 5th or early 6th century, have been discovered. These were situated on low terraces, and in each case the central structure was an octagonal wooden pagoda with sides 10 metres long. The pagoda was probably a tall, multistoried structure in the style of the Yingning Temple pagoda built in Lo-yang, China, in AD 467. Facing the central pagoda on three sides (north, east, and west) were Buddha chapels. This arrangement, one pagoda with three surrounding halls, seems to have been the earliest Buddhist temple plan used. The very same plan can be seen at a Paekche temple site in Puyŏ and at the site of the Asuka-dera temple near Kyōto, Japan. The Paekche temple also had a central octagonal wooden pagoda. In Silla, however, as can be seen in the well-known Hwang'yong Temple of Kyŏngju, the Koguryŏ-Paekche plan was modified to a one-pagoda (south), one-chapel (north) system.

Though the wooden structures of the period have been completely destroyed, three stone pagodas still exist, two in the Paekche area and one in Kyŏngju. At first Koreans built replicas of Chinese multistory wooden pagodas; but, since wooden structures were expensive and difficult to maintain, the idea arose, first in Paekche, of using stone. Paekche architects initially tried to copy the wooden pagoda as faithfully as possible. A good example of this is the stone pagoda at the Mirŭk Temple south of Puyŏ. Later, however, pagodas became smaller, and architectural details were much simplified, as can be seen in the five-story pagoda in Puyŏ. The square pagoda stands on the elevated platform of granite, and each story is capped by a thin roofstone with projecting eaves. The stories diminish progressively in size as they go upward, forming a characteristic slender and stabilized type from which the later Silla pagodas evolved. The only remaining Silla pagoda is at the Punhwang Temple in Kyŏngju, constructed in 634, a stone version of a Chinese brick pagoda of the T'ang dynasty (618–907).

Painting. Paintings from the Three Kingdoms are mainly those from decorated tombs. The earliest dated Koguryŏ tomb, the Tomb of Tongsu, or Tomb No. 3, in 'Anak, south of P'yŏngyang, was built in 357. All other known tombs except for Tokhŭng-ni Tomb, bearing an inscription datable to 408 AD, are undated but can be roughly classified as early (4th century), middle (5th century), or late (6th century). The early tomb murals were portraits of the dead master and his wife, painted either on the nichelike side walls of an entrance chamber or on the back wall of the main burial chamber. The paintings were executed in fresco, a technique of painting with water-soluble pigments on plaster. The colours used were black, deep yellow, brownish red, green, and purple. The general tone of the paintings is subdued and often gives a strange melancholic impression. In the middle stage, though portraits were still painted, they depicted the dead master in connection with some important event in his life, rather than seated solemnly and godlike as in the earlier period. In the Tomb of the Dancing Figures in the T'ung-kou region around Chi-an, the master is shown on the northern wall of the main chamber feasting with visiting Buddhist monks. A troupe of dancers is painted on the eastern wall and a hunting scene on the western one. The delicate wiry outlines of the first phase of Korean mural painting are replaced by bold, animated lines, which are quite distinct from the prevailing Chinese styles. In the hunting scene, mounted warriors shoot at fleeing tigers and deer. Lumps of striated clay are used to depict mountain ranges. Forceful brushstrokes are used to heighten the effect of motion of the galloping horses and fleeing game. This naive sense of dynamism is characteristic of Koguryŏ painting.

Tombs decorated with frescoes

In the third and final stage of Koguryŏ mural art, the technique of fresco painting was improved and imagery refined. Lines flow and colours are intensified. Genre paintings of preceding stages disappeared and the Four Deities of the cardinal compass points now occupied the four walls in Chinese fashion, a concept derived from Taoist religious art of the T'ang period. Dating probably from the first half of the 7th century, the paintings of the Three Tombs at Uhyŏn-ni, near P'yŏngyang, and of the Tomb of the Four Deities in Chi-an are the best examples from the final phase of Koguryŏ fresco painting.

Tomb painting spread to Paekche, where two examples of tomb wall painting can be found, the tombs of Songsan-ni in Kongju and of Nŭngsan-ni in Puyŏ. In addition, a pillow from the tomb of King Muryŏng (501–523), in Kongju, features fish and dragons and lotus flowers painted in flowing exquisite lines in ink against a red background. In the greater Silla area, one decorated tomb at Koryŏng in the former Kaya territory and two tombs discovered in the 1980s at Yŏngju have survived, but the paintings in all three are badly damaged. The best example of painting from the Old Silla period is found on a saddle mudguard made of multi-ply birch bark discovered in the Tomb of the Heavenly Horse in Kyŏngju in 1973; the mudguard depicts a galloping white horse surrounded by a band of floral design.

Beginnings of Buddhist sculpture

Sculpture. Buddhist sculpture probably began in the Koguryŏ kingdom in the 5th century. No 5th-century pieces survive, however, except for some fragments of terra-cotta figures. The earliest dated Koguryŏ Buddhist image is a gilt-bronze standing Buddha. It has an inscribed date that may correspond to the year 539. The elongated face, the flared drapery, and the mandorla or almond-shaped aureole, decorated with a flame pattern, all point to the influence of Chinese sculpture of the Northern Wei period (386–534.) A close adherence to the stylized linear tradition of northern Chinese sculpture was, in fact, the main characteristic of Koguryŏ sculpture.

In Paekche the Koguryŏ-type Buddha became more naturalistic and thus more Korean in style. The Buddha's face is rounder and more expressive, with the distinctive "Paekche smile." The style was apparently influenced by the softly modeled sculpture of southern China, particularly of the Southern Liang dynasty (502–557), when many Chinese artisans are believed to have come to Paekche. The best piece among some 18 or so extant Paekche gilt-bronze Buddhist images is a standing bodhisattva, or figure of one who has attained enlightenment. Now in the Ichida collection in Japan, it was originally from a temple site in Puyŏ. The seated Maitreya, or image of the future Buddha dwelling in the Tuṣita heaven (National Museum of Korea, Seoul), is of unknown provenance, although the round face, the well-proportioned feminine body, and the animated, naturalistic drapery suggest Paekche workmanship of about 600. The pinewood bodhisattva in the Kōryū Temple, Kyōto, Japan, has the same facial expression and posture, and it is believed to be the Maitreya sent from Korea in 623, as is recorded in *Nihon shoki,* the official history of Japan compiled in the 8th century. To-

ward the end of the Paekche dynasty, rock-cut sculpture, in the form of relief figures on exposed outdoor rocks, appeared. Dating from the mid-7th century, the first such example is at Sŏsan in South Ch'ungch'ŏng province. It is a Śākyamuni triad, or the historical Buddha flanked by the bodhisattvas Mañjuśrī and Samantabhadra.

Silla followed the naturalistic Paekche style but in a more static and conservative fashion. The seated gilt-bronze Silla Maitreya in the National Museum of Korea is of the same size as the Paekche Maitreya and is cast in the same pose of a half cross-legged figure in meditation. The drapery, however, is very conventionalized, and the image lacks the animation of the Paekche statue. In the 7th century the creation of stone sculpture increased in the Silla kingdom. Kyŏngju became the centre of production. Much of this stone statuary reflected influences from early T'ang sculpture of the 7th century, particularly in the characteristic interest in the body mass.

Decorative arts. Metalwork was one of the most developed mediums of the decorative arts in the Three Kingdoms period. Kings and high-ranking officials wore gold or gilt-bronze crowns and diadems and also adorned themselves with earrings, necklaces, bracelets, and finger rings made of gold, silver, bronze, jade, and glass. The best surviving pieces of jewelry and regalia come from intact Silla tombs. Only five gold crowns, coming from five Kyŏngju tombs, had been discovered by the early 1990s. The most elaborate, discovered in 1921 in the Tomb of the Golden Crown, consists of an outer circlet with five upright elements and a separate inner cap with a hornlike frontal ornament. It is made of cut sheet gold, and three of the frontal uprights are trees done in a highly stylized manner, flanked by two antler-shaped uprights. Numerous spangles and crescent-shaped pieces of jade (*magatama*) are attached to the vertical elements by means of twisted wire. The worship of trees and antlers was almost universal among ancient peoples of Central and Northern Asia, where the Koreans of the Three Kingdoms originated.

Silla crowns

The most representative type of Three Kingdoms pottery is the hard, grayish, unglazed stoneware of Silla. The predominant type of vessel forms are mounted cups and jars with erect cylindrical necks. At the foot of the cups are four or more rectangular apertures. There are also many human and animal figurines. In Paekche, tiles of gray clay were produced around Puyŏ in the 7th century, many with reliefs of boldly stylized landscapes.

Unified, or Great, Silla period (668–935). In 660 and 668, respectively, the Paekche and Koguryŏ kingdoms fell to the allied armies of the Silla king and the T'ang Chinese emperor, creating a new political and cultural era referred to as the Unified Silla period. This was the golden age of Korean art. Buddhism enjoyed a renewed prosperity, and great temples sprang up one after another in the Kyŏngsang region. Monks and scholars traveled to T'ang China to partake of its brilliant cosmopolitan culture. Korean culture was rapidly Sinicized. The capital city of Kyŏngju (like the contemporary Japanese capital of Heian-kyo, later Kyōto) was modeled after the T'ang capital of Ch'ang-an, with broad, straight avenues laid out on a rectangular grid pattern. From this time on, southern Korea, particularly the southeast, became the centre of Korean artistic development, and northern Korea, where once an aggressive Koguryŏ art had flourished, diminished in importance.

The Unified Silla period produced more granite Buddhist images and pagodas than any other period. Architectural ornamentation, such as roof tiles decorated with floral and animal designs, was of high quality. The bronzesmiths of Unified Silla did excellent work, as exemplified in numerous huge temple bells, *śarīra* boxes (containing sacred ashes of the Śākyamuni Buddha), and Buddhist statues. Toward the end of the reign, bronze seems to have been in short supply, and statues were cast in iron. One Buddhist painting has survived from the Unified Silla period. It depicts a Buddhist sermon held in a temple. Figures and architecture are represented in fine gold lines on blue-brown paper.

Architecture. Many temples were built during the Unified Silla period, and existing ones were enlarged. The low skyline of Kyŏngju must once have been dominated by towering pagodas. The original layout of a Unified Silla temple is best preserved in the Pulguk Temple to the southeast of Kyŏngju. The temple, constructed in the mid-8th century, is situated at the foot of a mountain and is divided into two adjoining complexes. The approach from the south is by a pair of stone staircases. The main complex is the eastern one, with two stone pagodas, one behind the entrance gate and the other in front of the main hall. One pagoda (Sŏkkat'ap) is in the typical square Silla style and represents Śākyamuni, the historical Buddha. The other pagoda (Tabot'ap) is more elaborate and symbolizes the Prabhūtaratna Buddha. The arrangement apparently symbolizes the Buddhist legend that, when Śākyamuni preached the *Avataṃsaka-sūtra,* the pagoda of Prabhūtaratna emerged out of the earth in witness of the greatness and truth of his preaching. A lecture hall once stood behind the main hall. A long, roofed corridor once surrounded the entire eastern complex, linking the main gate, main hall, and the lecture hall. The western complex symbolizes paradise and has a Hall of Paradise at its centre. The present wooden structures of the temple date from the 17th century, but the stoneworks, such as platforms, staircases, and foundation stones, all date from the Unified Silla period.

The Pulguk Temple

A special annex to the Pulguk Temple complex is the artificial cave temple, Sŏkkuram, on the crest of Mount T'oham about 1.6 kilometres away. The cave temple is a domed circular structure built of granite blocks and resembles a tholos, one of the beehive-shaped tombs built by the ancient Mycenaeans in Greece from about 1600 to 1300 BC. A square anteroom houses eight guardian figures in relief. The main chamber is 8 metres high and about 7 metres across. A large seated Śākyamuni (Amitābha, according to some) about 3.5 metres high, carved out of a single block of granite, occupies the centre on an elevated lotus pedestal. On the surrounding wall are 15 slabs in relief depicting 5 bodhisattvas and 10 disciples in attendance. The sculpture of this cave temple is without doubt the finest achievement of Buddhist art in the Orient.

The Paekche type of stone pagoda was adopted in the Unified Silla period, but certain architectural details of the earlier wooden pagoda were ignored and others simplified. The number of stories was reduced to three in most cases, and the main structure stands on a highly elevated, two-tier base. The roof stones have five-stepped corbels, or five projecting blocks supporting a superstructure. In the Śākyamuni pagoda at Pulguk Temple, a good example of the typical Unified Silla pagoda of the 8th century, the ratio of the widths of the stories from the bottom upward is 4:3:2, and the width of the lower base is equated with the height of the main structure above the upper base. This deliberate layout, unique to the Unified Silla period, makes the 8th-century pagoda a well-balanced and beautifully proportioned structure.

The Prabhūtaratna pagoda at Pulguk Temple is an exceptional case that demonstrates the skill of Unified Silla masonry. It is actually an enlarged stone version of an elaborate *śarīra* shrine. The main shrine, surrounded by railings, is supported by a rooflike square slab resting on four pillars with massive brackets, or supporting elements, to carry a projecting weight. The pillars in turn stand on an elevated platform approached by four staircases. On top of the octagonal main shrine is a small, similarly shaped roof adorned with a complex finial, or crowning ornamental architectural detail.

In addition to Buddhist architecture, other forms, including palatial architecture, flourished. Evidence of the magnificence of the Silla palace in Kyŏngju can be seen in the restored Anapchi (Goose and Duck Pond), a man-made pond originally constructed during the reign of King Munmu (661–681). When the pond was dredged in 1976, the original stone-built banks and a complex device for regulating the intake and outflow of water were discovered. Sites of pond-side pavilions as well as huge natural rocks that had been placed on the slopes of islets and on the banks of the pond also were uncovered, revealing the original layout of the 7th-century royal garden.

Sculpture. The sculpture of the Unified Silla period

Naturalistic granite images

was the high point of Korean naturalism and is marked by an abundance of statues in granite. During the first phase of the period, Korean sculpture was under the fresh influence of Chinese sculpture of the early T'ang period. Unified Silla works showed a certain vigour, though they were often stiff and had an imposing body mass. The tortoise base for the monument of King Muyŏl (d. 661) in Kyŏngju and a Śākyamuni triad at Kunwi are good examples of the first phase.

At the outset of the 8th century, however, Unified Silla sculpture freed itself of stiffness and took on a softened naturalistic look. The standing Amitābha and Maitreya (dated 721) from the site of Kamsan Temple may be considered typical examples of the first half of the 8th century and as stylistic stepping stones leading to the fully mature sculptures of the Sŏkkuram cave temple of the mid-8th century. The main Buddha of the cave temple has a massive body and a full, round face. Yet this is no mere hulking physical mass of monumental stone. The tranquil facial expression, the solid massive curves of the upper torso, and the somewhat formalized, simple drapery are skillfully synthesized and radiate the spiritual power and grace of the Buddha. In the case of the bodhisattvas, shapely feminine bodies are superbly reproduced on the rough granite surface; the curves, however, are covered by thin robes, executed in a stylized manner to de-emphasize the physical attractions and enhance the spiritual qualities. These figures may have been inspired by similar T'ang figures, such as those executed in 703 for the Pao-ching Temple in Sian, China. The Sŏkkuram figures, however, lack the secular and erotic character of the T'ang sculptures.



Saehan International

Monument at the tomb of King Muyŏl, 661, Silla period. In Kyŏngju, Kyŏngsang province.

Stylistic and technical degeneration, however, had already begun in the second half of the 8th century, as is indicated by the two seated bronze Buddhas in the Pulguk Temple. They retain the round, fleshy face of the Sŏkkuram Buddha, but their torsos are overly elongated and the drapery somewhat stylized, so that the spiritual quality is diminished. This mannered style of handling the image increases until the end of the century.

In the 9th century the Unified Silla kingdom itself began to decline. Sculptors were constrained to reduce the size of their pieces, both carved and cast. As a result statues were often out of proportion. A large square block representing the head might be placed on top of a small shrunken body with narrow, sloping shoulders. From about the mid-9th century, bronze came to be used only for small statuettes; large images were cast in iron.

Decorative arts. A considerable number of ceramic urns have been discovered, mainly in the vicinity of Kyŏngju. They are covered with stamped floral patterns, and some have a yellowish green lead glaze. The stamping and glazing were new techniques introduced by potters in the 7th century. Earthenware roof and square floor tiles also were produced. These were decorated with delicately molded lotus and other rich floral designs and were made for Buddhist temples and palace buildings.



Kyŏngju National Museum

Bronze bell of King Sŏngdŏk, 771, Unified Silla period. In the Kyŏngju National Museum. Height 3.33 m.

Bronze bells

Bronze work was outstanding in this period, especially the large bronze Buddhist bells. Four Unified Silla bells with inscribed dates survive, two of which are in Japan. A Korean bell of this period differs from a Chinese or Japanese example by the hollow cylindrical tube erected on the crown, alongside the traditional arched dragon handle, and in the surface decoration: the upper and the lower rims of the body are each surrounded by an ornamental horizontal band. Silla skill in casting is best seen in the colossal bronze bell of King Sŏngdŏk that was made in 771 for the Pongdŏk Temple and is now in the Kyŏngju National Museum.

Buddhist bronze shrines for *śarira* were sometimes placed inside stone pagodas. The best example is from the western pagoda of the Kamŭn Temple site. It is a square platform



By courtesy of the National Museum of Korea, Seoul

Śarira casket of the Kamŭn Temple, Unified Silla period (668–935). In the National Museum, Seoul. Height 15 cm.

on which a miniature glass bottle containing the *śarira* is placed under a rich canopy supported by four corner poles. The shrine was encased in a square outer box with a pyramidal cover, each panel of the box adorned with a bronze relief figure of one of the Four Guardians.

Koryŏ period (918–1392). In 935 the Unified Silla monarchy was supplanted by the newly risen Koryŏ dynasty (918–1392). Buddhism once again prospered under royal patronage. Koryŏ's close cultural ties with China during the Sung period (960–1279) resulted in direct influences from the advanced Chinese urban culture, and highly refined, Sinicized lifestyles prevailed among the aristocrats, the more important court officials, and the high-ranking Buddhist priests. The peace of the realm, however, was often disrupted by invaders from Manchuria, first Khitan, then Juchen, and finally by the Mongols. In 1232 the Koryŏ court fled to Kanghwa Island off the west coast of Korea, leaving the country to Mongol devastation and control. The art of Koryŏ never again equaled its pre-Mongol achievements.

Few original examples of Koryŏ architecture have survived. Koryŏ stone sculpture and stone pagoda construction were inferior to that of the Unified Silla period. Examples have survived largely because the Buddhist monks buried their images and ceremonial vessels before abandoning the temples to the Mongols. Good bronze temple bells were cast, although they were smaller in size than those produced in the Unified Silla kingdom. Buddhist sutras were painstakingly copied by monks in gold and silver on thick purple paper. Printing and wood-block engraving were innovations that reached a high state of development. A Koryŏ book is comparable in printing technique to the finest Chinese editions of the Sung period. The famous engraved edition of the entire *Tripiṭaka*, a long Buddhist canonical text, was done in Kanghwa Island in the mid-13th century as a commission of the government in exile. More than 80,000 engraved word blocks were used to print this edition. The major artistic achievement of the Koryŏ period was the production of porcelain with a celadon glaze. Sets of celadon ware were customarily buried with the dead, and it is from these tombs that most of the Koryŏ celadon available in the 20th century has come.

Architecture. Traditional Korean architecture must have been similar to T'ang architecture, which is best illustrated by the main hall of Nan-ch'an Temple (782), Shansi, China. The main hall of the Tōshōdai Temple in Kyōto, Japan, also is believed to be a good example of T'ang-style architecture. In Korea the adaptation of the T'ang architecture is called the *chusimp'o* style. It is characterized by the so-called column-head bracketing, or complexes of brackets that project above the heads or capitals of the columns, with or without intercolumnar struts (inclined supports). One of the best examples of *chusimp'o* architecture is the Muryangsujŏn ("Hall of Eternal Life") of Pusŏk Temple. Dating from the 13th century, this is believed to be one of the oldest wooden structures in Korea.

The *chusimp'o* and *tap'o* styles

About 1300 a new architectural style was brought in from Sung China. Called *tap'o* (multi-bracket), it is characterized by intercolumnar bracketing in place of struts. *Tap'o* became the main style during the following Chosŏn dynasty. Built in the *tap'o* manner are the Pokwangjŏn hall of the Simwŏn Temple and the Eungjinjŏn hall of the Sŏkwang Temple, both of which are datable to the second half of the 14th century. The new *tap'o* buildings are much more decorative than those in the *chusimp'o* style because the intercolumnar brackets fill up the otherwise empty spaces between columns.

The early Koryŏ pagoda was executed in the Unified Silla style, although the roof stones were thinner and the number of eave corbels decreased to three or four. Then, after a short period, this style of pagoda changed drastically. The number of stories increased, the corbels on the roof stones became almost unrecognizable, and the height of each story was reduced. The Koryŏ pagoda became either an emasculated pagoda of the Unified Silla period or an unstable columnar silhouette. There are also towering octagonal pagodas as, for example, the one at Wŏlchŏng Temple. These were not revivals of the Koguryŏ type but a contemporary style imported from Sung China.

Toward the end of the Koryŏ the building of pagodas virtually came to a halt. One exception is the 10-story (12-metre) marble pagoda built in 1348 for the Wŏngak Temple in Kaesŏng (now in the Kyŏngbok Palace, Seoul). The pagoda stands on a cross-shaped, three-tiered platform. Every architectural detail from roof tiles to the bracket system is painstakingly reproduced, and numerous Buddhist figures in relief cover the entire surface of the pagoda.

Painting. Only about 10 examples of original Koryŏ painting are extant, and most of these are in Japan. They are mainly minor works on Buddhist themes except for several badly worn fragments of a hunting scene attributed to King Kongmin (1351–74) and two landscapes by other artists. There is little to be said about these isolated works except that they are in varying degrees in the style of Chinese painting of the Sung period (960–1279). Among the few examples of Koryŏ temple wall paintings are the Buddhistic images in the Chosa-dang (Founder's Hall) at Pusŏk Temple (1377) and the paintings of flowers in the Main Hall of the Sudŏk Temple (1308). Among the important examples of Koryŏ tomb painting is an image of a flying *deva* (from the 12th or 13th century) discovered in 1971 on the wall of a tomb at Kŏch'ang in southeastern Korea.

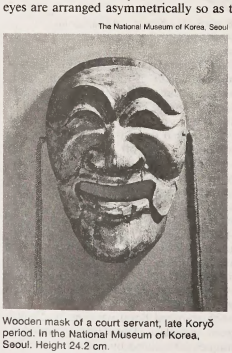
Sculpture. Compared with that of the Unified Silla period, Koryŏ sculpture shows a decline in both quantity and quality. However, before the decline a momentary surge of naturalism, a traditionally northern Korean quality, revitalized the period. Large images with imposing bodies and archaic smiles were successfully cast in iron, a medium not used since the late Unified Silla period. Direct copies from 8th-century Unified Silla models were often attempted. The colossal seated iron Buddha in the National Museum of Korea is the best example of this revival style. This image of the Buddha was clearly influenced by the large Śākyamuni of the Unified Silla cave temple of Sŏkkuram. Only the long narrow eyes, the sharpened nose, and a certain angularity in the treatment of the drapery give the Buddha a unique Koryŏ coldness that heralds the rather abstract quality found in later iron images.

In stone sculpture, also, the revival style is noticeable. The trend, however, was short-lived, and by the 12th century Koryŏ sculptors seem to have lost the art of working large, fully rounded figures in stone or metal. The decline in technique was manifested in the abstract tendency of certain figures of the middle of the Koryŏ period, such as the seated iron Buddha in Ch'ungju.

Although the sculpture produced by the major workshops suffered a decline, good sculptors could still be found in the countryside. One of the best known is the master who carved a set of wooden play masks for the village of Hahoe near Andong in southeastern Korea. The masks are marked by an exotic realism. The deep-set eyes are arranged asymmetrically so as to become mobile

Theatrical masks

The National Museum of Korea, Seoul



Wooden mask of a court servant, late Koryŏ period. In the National Museum of Korea, Seoul. Height 24.2 cm

under the play of changing light and shade. The nose, very un-Korean, is extraordinarily long and aquiline. The separately made chin, like the nose, is massive. Models for these exotic masks must have come from China, as early as the T'ang dynasty, when elements of Persian and Central Asian art found their way into China. These Korean masks might well have served as the intermediary links through which the Japanese mask for the nō drama developed from original Chinese models.

Decorative arts. Traditional Koryŏ pottery was unglazed grayish stoneware in the Unified Silla tradition. By the end of the 10th century, however, the technique of high-fired, green-glazed porcelain of the Yüeh type was introduced from Chekiang province in southern China. After an initial period of imitation, Koryŏ potters, from about the mid-11th century or slightly earlier, started to produce their own distinctive kind of porcelain with a celadon glaze. Two main ceramic centres, at Kangjin and Puan, operated in southwestern Korea from the very beginning to the end of the Koryŏ period.

Celadon porcelain

The first period of Koryŏ celadon, from about 1050 to 1150, was the period of plain celadon ware. The "secret" colour of Koryŏ celadon, a greenish blue with a mysteriously deep tone, was regarded by the Sung Chinese as one of the "ten best things in the world." The potters of the first period appear to have been mainly concerned with the deep, lustrous colour and the formal beauty of the vessel, although they also used incised, engraved, or molded animal and floral patterns to decorate their vessels. Their specialties were animal- and fruit-shaped ewers and incense burners. White procelain of the Chinese *ying-ching* type also was produced during this period, though only in limited quantities.

The next 100 years, from 1150 to 1250, is the period of inlaid celadon ware. The technique of inlay on celadon is generally believed to have been invented about the mid-12th century. The idea of inlay may have come from a number of sources, but it is undoubtedly related to techniques of metal inlay that in turn were derived from inlaid lacquer. Whatever the origin, inlaid celadon was a Korean invention and unique to the Korean pottery of the 12th to the 15th century. In this technique, the freshly thrown vessel is left to dry to a leatherlike hardness. Designs are then incised or gouged out and filled with white or black clay. Sometimes, instead of the design, the background is scraped off and filled with black or white clay. During the initial stage, potters were still aware of the importance of glaze colour, despite the remarkable effect of inlaid designs. As time passed, however, they gradually inclined toward the decorative effect of designs, and the space occupied by the design came to dominate their work. The famous vase in the Kansong Art Museum, Seoul, is an outstanding example of this mature period of inlaid celadon.

From about 1250 to the end of the Koryŏ period in 1392 is the period of decline. The inlay technique continued, but the designs were loose and coarse and lacked the craftsmanship of the earlier pieces. The glaze colour is predominantly yellowish because an oxidizing fire was used. Crowded floral patterns painted in an iron type of underglaze became fashionable under the influence of Chinese pottery of the Yüan period (1206–1368).

The Koreans probably learned the technique of lacquer making from the Chinese at Nangnang during the early years of the Three Kingdoms period. It thenceforth became so popular that inlaid lacquer is almost completely a Korean specialty. The technique, although called "inlaid," is more accurately a polish-expose technique. Cut pieces of abalone or tortoiseshell, supplemented by silver or bronze wire, are pasted on the hemp or hemp-coated pinewood core with a thick coat of lacquer. Many layers of lacquer and special glue are then applied to the design until the shell layer is completely concealed. It is then polished with whetstone and charcoal until the surface of the design is revealed.

Inlaid lacquerware

Bronze temple bells continued to be cast, but they gradually were reduced in size, and the craftsmanship showed a remarkable decline from the Unified Silla period. A Koryŏ bell is distinguished by the outer edge of the crown, which characteristically is marked by a band of lotus petals that projects out obliquely. Images of outlined Buddhas and bodhisattvas around the trunk replaced the earlier flying *devas* (heavenly beings who are the guardians of Buddhism).

Important among the Koryŏ bronzes is the series of beautifully finished incense burners still treasured by many temples. These censers look like enlarged mounted cups with deep bowllike bodies, the mouth rims of which flare out horizontally to form a broad brim. The body is mounted on top of a conical stand with graceful concave side lines. The surface of the vessel is always covered with fluent, linear floral patterns or animated dragons inlaid with silver, which stand out strikingly against the shining black patinated background. The same techniques and decorative motifs also were used for making the artistically outstanding bronze mirrors typical of the Koryŏ period.

Chosŏn (Yi) period (1392–1910). In 1388 General Yi Sŏng-gye dethroned the pro-Mongol King Wu. Four years later, in 1392, General Yi proclaimed himself founder of the new Chosŏn dynasty (1392–1910) and moved the capital from Kaesŏng (Songdo) to Seoul. His policy was to maintain close political and cultural ties with Ming China (1368–1644). Buddhism, by then thoroughly corrupt, was displaced as the state religion by a puritanic Neo-Confucianism, then also on the ascendant in Ming China. Confucianism became the dominant influence on Korean thought, morals, and aesthetic standards. Chosŏn craftsmen and artisans, unable except occasionally to draw inspiration from imported Chinese art, were constrained to rely upon their own sense of beauty and perfection. Chosŏn artists, particularly in the decorative arts, showed a more spontaneous, indigenous aesthetic sense than the sophisticated aristocratic elegance of Koryŏ art.

In 1592 the Japanese general Toyotomi Hideyoshi invaded Korea. For many years the entire peninsula was a battlefield, and a tremendous amount of art was destroyed. The Japanese even carried off many Korean potters, who later managed to settle in the northern part of the island of Kyushu and become the founders of the Japanese porcelain industry. The Japanese invasion was soon followed by the Manchu, a Manchurian people, who later conquered China and established the Ch'ing dynasty (1644–1911/12). The two invasions left the Chosŏn government in a critically weakened condition, but they also inspired the rise of a strong nationalist sentiment among the Korean people. Concern focused on solving domestic social problems and on reviving and restoring confidence in Korean culture and identity. Scholars made efforts to develop practical knowledge and wisdom to improve life in Korea rather than studying "empty" Confucian theories. Painters for the first time showed profound interest in the landscape and daily life of Korea, and Chosŏn art of the 17th to 18th century demonstrated a marked Korean character and flavour. This florescence of Chosŏn art ended after only two centuries, however, because of the lack of public and private patronage, the lack of inspiration, and the apathy and poverty that occurred as the dynasty itself entered the last phase of its history.

Nevertheless, this period left abundant artistic remains. There are many palace and temple buildings, although few date to before the Japanese invasion. Buddhist images were usually made of wood instead of bronze and iron, and granite was rarely used for sculpture. Among the secular arts, painting and ceramics were the most important. The Chosŏn government maintained an Office of Painting, or Imperial academy of painting (Tohua-sŏ), and the government also operated an official kiln that alone was authorized to produce blue-and-white porcelain. Local private kilns also mass-produced large amounts of ceramics. The Chosŏn dynasty was finally terminated when Japan annexed Korea in 1910.

Architecture. Many large palace and temple buildings are preserved from the Chosŏn period, particularly those built from the 17th century on. The *chusimp'o,* the columnhead bracket style of the Koryŏ period, continued during the early part of this period. But the dominant architectural style of the Chosŏn period was the *tap'o,* or the intercolumnar bracket style. At least five large palaces in Seoul alone date from the beginning of the period. The

Kyŏngbok Palace

largest and most important is the Kyŏngbok Palace, originally a complex of more than 100 buildings. The entire palace was burned down during the Japanese invasion in the late 1500s but it was reconstructed between 1865 and 1867. Kŭnjŏng-jŏn, the palace's throne hall, built in the decorative *tap'o* style, is the largest wooden building in Korea. Among temples, the main halls of Muwi Temple, Kaisim Temple, and Pongjong Temple belong to the early Chosŏn period, while the grand main hall of the Hua'ŏm Temple represents the later Chosŏn *tap'o* architecture. The only important Chosŏn pagoda is the marble pagoda (1467) in Seoul's Pagoda Park.

Painting. Chosŏn painting up to the end of the 16th century was dominated by court painters attached to the Office of Painting. Their style followed that of Chinese professional court painters, the so-called Northern school of Chinese painting, and was thus variably influenced by the Kuo Hsi school of the Northern Sung, the Ma-Hsia school of the Southern Sung, and the Che school of Ming China. Famous painters of the period are An Kyŏn, Ch'oe Kyŏng, and Yi Sang-cha. An Kyŏn's best work, "Dream Journey to the Peach Blossom Land" (1447; Tenri University Collection, Japan), executed in the heroic style of the Northern Sung, is a horizontal scroll depicting fantastic mountains and streams dotted with peach blossoms.

Yi Am, Shin Sa-im-dang, and Yi Chŏng are the better scholar-painters of the first period. Unlike the professional court painters, who made Chinese landscapes their specialty, these amateur scholar-painters devoted themselves to painting the so-called Four Gentlemen—the pine tree, bamboo, plum tree, and orchid—as well as such traditionally popular subjects as birds, insects, flowers, and animals.

In the early 17th century the Southern school of China, exemplified by Mi Fu, Shen Chou, Wen Cheng-ming, and others, strongly influenced Korean painters, particularly the nonprofessional scholar-painter. Professional academic painters followed the academic court style of Ch'ing China, which was itself a sort of formalized Southern style. The "expressionistic" and individualistic Ch'ing style of the Eight Eccentrics of Yang-chou, however, did not find followers in Chosŏn Korea. Concurrent with the imitation of Chinese painting styles was a movement to achieve in approach and effect a truly Korean expression. The works of Cho Sok, noted for their thin wash of ink, displayed the melancholy of Chosŏn society. Chŏng Sŏn, a great Chosŏn master, disliked the imaginary Chinese-style landscapes and devoted himself instead to the real Korean landscape. His favourite theme was the rugged peaks of Mount Kumgang (Diamond Mountain) in central eastern Korea. To depict rocky cliffs and soaring forests, he devised his characteristic "wrinkles" of forceful vertical lines. The trend toward a national style, established by Chŏng Sŏn and others, was followed by Kim Hong-do, Shin Yun-bok, and Kim Tŭk-shin, who all painted national scenes of daily life in Korea with a realism that often bordered on caricature. Among these the greatest master was Kim Hong-do, better known under the name of Tanwŏn. He also painted many Korean landscapes and was the first Korean painter to draw his genre themes from the life of the lower classes. He seems also to have been the first Korean painter to try to depict human muscles.

Development of a Korean style

In the 19th century, Cho Chŏng-kyu, Hŏ Yu, Chang Sŭng-ŏp, and Cho Sŏk-chin were among the more active professional painters. Their paintings were mannered and exhibited an academic style lacking individuality. They painted many excellent portraits of Korean dignitaries in an indigenous style, but otherwise they returned to the old clichés of pseudo-Chinese painting. During this century the first influence of European art on the court painters may be seen in their use of shading techniques in painting portraits.

The activities of a short-lived group of painters who followed the *wen-jen hua*, or Chinese literati style of painting, should be seen against the general decline of the academic style of the 19th century. All of them were men of learning and genuine taste who grasped the spirit of such great Chinese masters of the Yüan period as Ni Tsan and Huang Kung-wang. The most distinguished members of this group were Kim Chŏng-hŭi, the great calligrapher, who painted little, and Chŏn Ki, who died young.

Sculpture. By the beginning of the Chosŏn period, the production of traditional religious sculpture had virtually died out because of the importance of Confucianism, the new state religion. The Buddhist images that were produced are mostly made of wood and are artistically undistinguished. At first Chosŏn sculpture followed the late Koryŏ style and early Ming sculpture of the late 14th and 15th centuries. A Chosŏn Buddha is characterized by a rounded late Unified-Silla-type head with a flat, emotionless face. The body is a simple, stolid mass covered with a loose, yet leatherlike, thick robe. Drapery folds are depicted in a formalized, schematic series of plaits.

Decline of Buddhist sculpture

Secular sculpture included the series of stone statues of civil and military officers that were erected in front of the tombs of members of the royal family and other dignitaries.

Decorative arts. Although a wide variety of decorative arts flourished in the Chosŏn period, the making of pottery and porcelain was especially important. One of the most popular types of ceramic ware produced was called *punch'ŏng*, the Korean term for a type of pottery known in Japan as *mishima*. This is a simplified form of *punjang ch'ŏngja*, or slip-decorated celadon. The slip-decoration includes inlaid, incised, and stamped patterns filled with white clay, and also the overall application of a white coating under the celadon glaze. Incision and painting in underglaze iron also are applied at times over the white coating. The technique evolved (or degenerated) from Koryŏ inlaid celadon, which had become coarse and rough in its final stages. The early Chosŏn potters invented a new device to produce the inlay effect more quickly and easily. A wooden or clay stamp with tiny embossed dots was used to produce designs of closely spaced depressed dots over the entire surface of a vessel in a matter of minutes. White clay was then rubbed into the dots and the excess clay wiped off. There are, however, Chosŏn pieces done in the traditional inlay technique, and they can be instantly distinguished from late Koryŏ wares by their crude and unsophisticated designs (floral as well as animal) and the stained grayish green colour of their glaze.

Punch'ŏng ware

The predominant *punch'ŏng* shapes are small or medium-size wine bottles and tea and rice bowls. Many were produced under orders from government offices, but their mass production suggests that there may have been increasing demand from the general public. *Punch'ŏng* pottery was loved by Japanese masters of the tea ceremony. The Hideyoshi invasion put an end to the lingering Koryŏ inlaid celadon once and for all. The *punch'ŏng* stamping technique, however, is still used on the island of Okinawa, south of Japan.

White porcelain, which may have been inspired by the Yüan and Ming blue-and-white porcelain ware of China, has remained as the most practical ware for ordinary Koreans. White porcelain wares of the pre-16th-century Chosŏn dynasty are covered with a milky-white devitrified glaze. They were produced at hundreds of central and local kilns, but the best pieces came from the Kwangju kilns south of Seoul during the 15th century. Besides being the

White porcelain

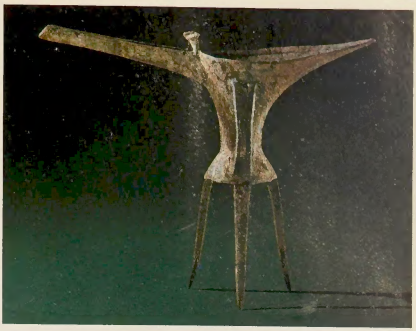


Punch'ŏng ware wine bottle in 'rice-bale' shape, 15th century, Chosŏn dynasty. In the collection of the Honolulu Academy of Arts. Length 26 cm.

By courtesy of the Collection of the Honolulu Academy of Arts



Ceremonial *ts'ung* of jade (calcined nephrite), 3rd millennium BC, Neolithic Liang-chu culture. In the Seattle Art Museum, Washington, U.S.



Ceremonial bronze *chüeh* (pre-Style I) from Erh-li-t'ou, Honan province, early 2nd millennium BC, pre-Shang or Shang dynasty. In the Cultural Relics Office, Yen-shih county, Honan province. Height 25.6 cm.

Ancient Chinese Art



Painted pottery funerary urn, Neolithic Pan-shan phase, from Yang-shao, Honan province, *c.* 3000 BC. In the Museum of Far Eastern Antiquities, Stockholm. Height 33.5 cm.



Ceremonial bronze *ho* pouring vessel (Style V), from An-yang, Honan province, *c.* 11th century BC, Shang dynasty. In the Nezu Institute of Fine Arts, Tokyo. Height 63 cm.



Ceremonial ivory goblet, inlaid with turquoise, from the tomb of Lady Fu Hao, An-yang, Honan province, *c.* 12th century BC, Shang dynasty. In the Archaeology Institute, Peking. Height 30.5 cm.



Ceremonial bronze *fang-tsun* (Style IV) from Ning-hsiang-hsien, Hunan province, *c.* 12th century BC, Shang dynasty. In the Museum of Chinese History, Peking. Height 58.3 cm.

Plate 1: (Top left) The Seattle Art Museum, purchased by the Foster family in memory of Albert O. Foster, photograph by Paul Macapia; (top right and bottom left) The Metropolitan Museum of Art, New York City, on loan from the People's Republic of China, photographs by Seth Joel; (centre left) Ostasiatiska Museet, Stockholm; (centre right) Nezu Institute of Fine Arts, Tokyo; (bottom right) Wang Lu/ChinaStock Photo Library



Painted limestone statue of Kuan-yin from Ch'ang-an (Sian), *c.* 570 AD, probably Northern Chou dynasty. In the Museum of Fine Arts, Boston. Height 2.49 m.



"Admonitions of the Court Instructress," detail of a hand scroll attributed to Ku K'ai-chih (*c.* 344–*c.* 406), possibly a T'ang dynasty copy of an Eastern Chin dynasty original. Ink and colours on silk. In the British Museum. Height 24.8 cm.

Ceremonial bronze *tsun* vessel in the form of a rhinoceros, originally inlaid with precious metals, from Hsing-p'ing Hsien, Shensi province, late 3rd century BC, late Chou–early Han dynasty. In the Museum of Chinese History, Peking. Height 34.1 cm.



Chinese art from the Chou dynasty through the Six Dynasties period

Bronze mirror back inlaid with gold and silver from Chin-ts'un near Lo-yang, Honan province, *c.* 3rd century BC, Chou dynasty. In the Eisei Bunko Foundation, Tokyo. Diameter 17.5 cm.



Bronze hill-censer (*po-shan hsiang-lu*), inlaid with gold, from the tomb of Liu Sheng, Prince Ching of Chung-shan, Man-ch'eng, Hopeh province, late 2nd century BC, Western Han dynasty. In the Hopeh Provincial Museum, Shih-chia-chuang. Height 26 cm.



Plate 2: (Top left) Francis Bartlett Fund, courtesy, Museum of Fine Arts, Boston; (top right) copyright British Museum; (centre right and bottom left) The Metropolitan Museum of Art, New York City, on loan from the People's Republic of China; photographs by Seth Joel; (bottom right) Eisei Bunko Foundation, Tokyo



Funerary banner from the tomb of Lady Tai (Hsin Chui), Ma-wang-tui, Hunan province, *c.* 168 BC or shortly after, Western Han dynasty. Ink and colours on silk. In the Hunan Provincial Museum, Ch'ang-sha. Height 2.05 m.



Bronze "Horse and Swallow" from the tomb of General Chang, Lei-t'ai, Wu-wei county, Kansu province, 2nd century AD, Eastern Han dynasty. In the Kansu Provincial Museum, Lan-chou. Height 32.4 cm.



Gilt-bronze Prabhūtaratna and Śākyamuni, 518 AD, Northern Wei dynasty. In the Guimet Museum, Paris. Height 26 cm.



Bodhisattva, detail of a painted mural from cave 272, Tun-huang, Kansu province, mid-5th century, Northern Wei dynasty.



Embroidered silk with dragon, phoenix, and tiger pattern, from Ma-shan Tomb No.1, 4th–3rd century BC, Chou dynasty. In the Ching-chou Museum, Hupeh province.

Plate 3: (Top left) Wang Lu/ChinaStock Photo Library, (top right) Robert Harding Picture Library, (centre right) Cliche des Musees Nationaux, Paris, © Photo R.M.N., (bottom right) Kong Lihang/ChinaStock Photo Library, (bottom left) Chen Zhi'an/ChinaStock Photo Library



Main hall of Nan-ch'an Temple at Wu-T'ai in Shansi province. 782 or earlier, T'ang dynasty; reconstructed 1974–75.



Ewer, three-colour glazed stoneware with dragon-head handles, 8th century, T'ang dynasty. In the Tokyo National Museum. Height 47.4 cm.

Chinese art from the T'ang and Northern Sung dynasties



"Travelers Among Mountains and Streams," hanging scroll by Fan K'uan (*c.* 960–*c.* 1030), Northern Sung dynasty. Ink and slight colour on silk. In the National Palace Museum, Taipei. 1.55 m x 74.3 cm.

Red sandalwood lute inlaid with mother-of-pearl decor, 8th century, T'ang dynasty. In the Shoso-in Treasure House, Nara, Japan. Length 1.004 m.



"Ming-huang's Journey to Shu," hanging scroll attributed to Li Chao-tao (fl. *c.* 700–730), T'ang dynasty style, possibly a 10th–11th-century copy of an 8th-century original. Ink and colours on silk. In the National Palace Museum, Taipei. 55.9 x 81 cm.



Plate 4: (Top left) Christopher Liu/ChinaStock Photo Library; (top right and bottom left) National Palace Museum, Taipei, Taiwan, Republic of China; (centre left) Collection of the Tokyo National Museum, gift of Dr. Yokogawa Tamisuke; (bottom right) Office of the Shoso-in Treasure House, photograph, Benrido Co. Ltd.



Buddhist guardian deity, three-colour painted ceramic sculpture from Chung-pao-ts'un near Sian, Shensi province, 8th century, T'ang dynasty. In the Shensi Provincial Museum, Sian. Height 65 cm.



Timber pagoda of the Fo-kung Temple at Ying-hsien, Shansi province; 1056, Sung dynasty.



"A Pheasant and Sparrows Among Rocks and Shrubs," hanging scroll attributed to Huang Chü-ts'ai, 10th century, Northern Sung dynasty. Ink and colours on silk. In the National Palace Museum, Taipei. 99 x 53.6 cm.



Kuan-yin and attendant bodhisattvas, detail of a painted mural from cave 57, Tun-huang, Kansu province, early 8th century, T'ang dynasty.

Plate fit: (Top left) Wang Lu/ChinaStock Photo Library; (top right) Christopher Lu/ChinaStock Photo Library; (bottom left) National Palace Museum, Taipei, Taiwan, Republic of China; (bottom right) Chen Zhi an/ChinaStock Photo Library



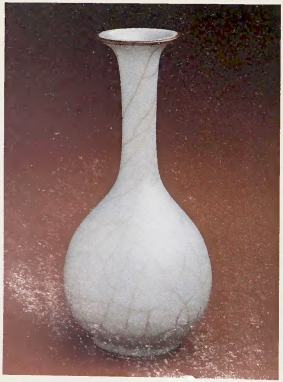
"Early Spring," detail of a hanging scroll by Kuo Hsi, dated 1072, Northern Sung dynasty. Ink and slight colour on silk. In the National Palace Museum, Taipei. Complete scroll 1.58 x 1.079 m.



Vase in the shape of a ceremonial *ts'ung, kuan* glazed stoneware, probably from Lung-ch'üan, Chekiang province, 12th–13th century, Southern Sung dynasty. In the Tokyo National Museum. Height 19.7 cm.

Chinese art of the Sung and Yüan dynasties

Glazed porcelain *kuan* ware vase, 13th century, Southern Sung dynasty. In the Percival David Foundation of Chinese Art, London. Height 18 cm.



"Five-Coloured Parakeet on Blossoming Apricot Tree," detail of a hand scroll by Emperor Hui-tsung, early 12th century, Northern Sung dynasty. Ink and colour on silk. In the Museum of Fine Arts, Boston. 53.3 cm x 1.251 m.



Plate 6: (Top left) National Palace Museum, Taipei, Taiwan, Republic of China; (top right) Collection of the Tokyo National Museum, gift of Mr. Hirota Matsushige; (bottom left) Percival David Foundation of Chinese Art, London; (bottom right) Mura Antoinette Evans Fund, courtesy, Museum of Fine Arts, Boston.



Ching-te-chen blue-and-white ware porcelain temple vase, dated 1351, Yüan dynasty. In the Percival David Foundation of Chinese Art, London. Height 63.4 cm.



"Fragrant Springtime, Clearing After Rain," album leaf by Ma Lin (d. after 1246). Southern Sung dynasty. Ink and slight colour on silk. In the National Palace Museum, Taipei. 27.5 x 41.6 cm.



Tea bowl, Chien-type stoneware with "sparkling" oil spots (*yohen temmoku*) from Fukien province, 12th–13th century, Southern Sung dynasty. In the Seikado Bunko Art Museum, Tokyo. Diameter 12.2 cm.

"Nine Horses," detail of a hand scroll by Jen Jen-fa, 1324, Yüan dynasty. Ink and colours on silk. In the Nelson-Atkins Museum of Art, Kansas City, Missouri. Complete scroll 31.2 cm x 2.62 m.



Plate 7. (Top left) Percival David Foundation of Chinese Art, London; (top right) National Palace Museum, Taipei, Taiwan, Republic of China; (centre right) © The Seikado Bunko Art Museum; (bottom) The Nelson-Atkins Museum of Art, Kansas City, Missouri (Purchase: Nelson Trust) 72-8



"Rats and Lamp," hanging scroll by Ch'i Pai-shih, 1947. Ink and colours on paper. In a private collection. 1 m x 33.6 cm.



Garden of the Master of Nets (Wang-shih Yüan) at Su-chou, Kiangsu province, Ming and Ch'ing dynasties.

Chinese art from the Ming dynasty through the modern period

The Forbidden City, Peking, Ming and Ch'ing dynasties.



Plate 8: (Top left) The Metropolitan Museum of Art, New York, The Harmsen Gift of Robert Hatfield Ellsworth; (top right) Caroline Courtauld; (bottom) photograph, Palace Museum, Beijing/Wan-go Weng Inc. Archive



"Boat People," hanging scroll by Ho Huai-shuo, 1979. Ink and colours on paper. In The Water, Pine and Stone Retreat Collection, Hong Kong. 66 x 66.5 cm.



Dome-shaped I-hsing ware teapot with a six-lobed body, signed Kung-ch'un, dated 1513, Ming dynasty. In the Hong Kong Museum of Art. Height 9.9 cm.



"Self Portrait," hanging scroll by Jen Hsiung (1820–57), ink and colour on paper. In the Palace Museum, Peking. 1.8 m x 78.8 cm.



"A Tall Pine and Taoist Immortal," hanging scroll with self-portrait (bottom centre) by Ch'en Hung-shou, dated 1635, Ming dynasty. Ink and colours on silk. In the National Palace Museum, Taipei. 2 m x 98.7 cm.

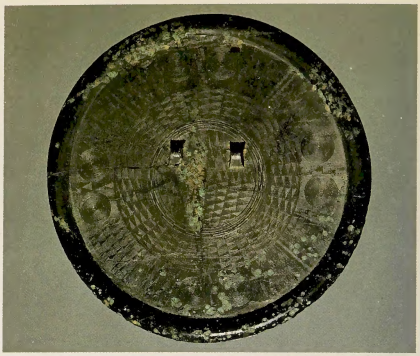
"Chairman Mao at Jinggang Mountain," oil on canvas by Luo Gongliu, 1961. In the Museum of Chinese Revolutionary History, Peking. 1.5 x 2 m.



Plate 9. (Top left) The Water, Pine and Stone Retreat Collection, Hong Kong. (top right) National Palace Museum, Taipei, Taiwan, Republic of China. (centre left) reproduction by permission of the Urban Council Hong Kong from the collection of the Hong Kong Museum of Art. (bottom left) Hu Chui/ChinaStock Photo Library. (bottom right) Zhao Liye/ChinaStock Photo Library



Gilt bronze seated Maitreya, *c.* 600 AD, probably Paekche. In the National Museum of Korea, Seoul. Height 93.5 cm.



Bronze mirror, *c.* 300 BC, Early Iron Age. In the Korean Christian Museum at Soongsil University, Seoul. Diameter 21.2 cm.

Korean art of the Early Iron Age and the Three Kingdoms period



Clay vessel in the shape of a mounted horseman, 5th–6th century, Silla period. In the National Museum of Korea, Seoul. Height 25.8 cm.



Mudguard with painted heavenly horse, colours on birch bark, *c.* AD 500, Silla period, from the Tomb of the Heavenly Horse, Kyŏngju, South Korea. In the National Museum of Korea, Seoul. 50 x 72 cm.

Plate 10: (Top left, bottom) National Museum of Korea, Seoul, (top right) The Korean Christian Museum at Soongsil University, Seoul, (centre right) National Museum of Korea; photograph, Siehan Color



Five-story stone pagoda in Puyŏ, South Korea, first half of 7th century, Paekche period. Height 8.33 m.



Gold crown, *c.* AD 500, Silla period, from the North Mound of Tomb 98, or the Great Tomb at Hwangnam-dong, Kyŏngju, South Korea. In the Kyŏngju National Museum. Height 27.5 cm.

Hunting scene from Tomb of the Dancing Figures, detail of a wall painting on plaster from near Chi-an, China, 5th–6th century, Koguryŏ period. In the National Museum of Korea, Seoul. Entire painting 3.1 x 1.5 m.



Plate 11: (Top left) Grafica Co., Inc., (top right) Kyongju National Museum, (bottom) National Museum of Korea, Seoul; photograph, Saehan Color



Pulguk Temple, near Kyŏngju, South Korea, 8th century.

Korean art of the Unified Silla period



Tile with *posanghwa* (floral) designs. *c.* 7th–8th century. In the National Museum of Korea, Seoul. 31.5 x 31.7 x 6.9 cm.

Tabot'ap pagoda from the Pulguk Temple, near Kyŏngju, South Korea, 8th century.



Seated granite Śākyamuni of the Sŏkkuram Grotto Shrine, Kyŏngju, South Korea, *c.* mid-8th century. Height 4.8 m.

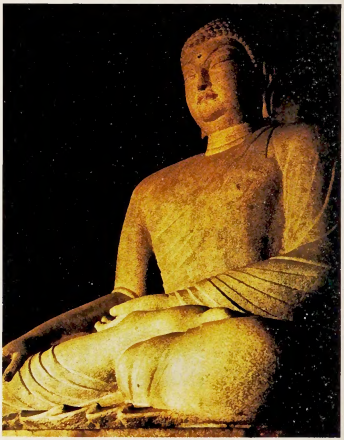


Plate 12: (Top, bottom left) National Museum of Korea; (bottom right) Korean Overseas Information Service. Color, (centre) National Museum of Korea, Seoul.

Muryangsujŏn hall of Pusŏk Temple, Andong, South Korea. wood, 13th century.





Lacquer box inlaid with mother-of-pearl, 12th century. In the Museum of East Asian Art, Cologne. Length 28 cm.

Bronze incense burner inlaid with silver. 1177. In the collection of the Pyochung Temple, Kyong Nam, South Korea. Height 27.5 cm.



Korean art of the Koryŏ period

Celadon vase inlaid with cloud and crane design, *c.* 13th century. In the Kansong Art Museum, Seoul. Height 42 cm.



Celadon wine cup with stand, first half of the 12th century. In the National Museum of Korea, Seoul. Height of cup 4.8 cm, height of stand 4.4 cm.

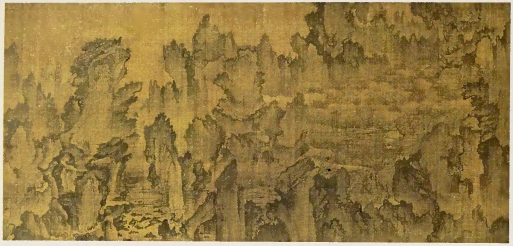


Plate 13: (Top left) Grafica Co., Inc., (centre right) Pyochung Temple, (bottom right) National Museum of Korea, Seoul, (bottom left) Kansong Art Museum, (centre left) Museum für Ostasiatische Kunst, Cologne



Kŭnjŏng-jŏn, the throne hall of Kyŏngbok Palace, Seoul, 1867.

"Dream Journey to the Peach Blossom Land," slight colours on silk by An Kyŏn, 1447. In the Tenri Central Library, Tenri University, Nara, Japan. 38.7 cm x 1.065 m.





Porcelain blue-and-white faceted bottle with floral design, 17th century. In the National Museum of Korea, Seoul. Height 41.5 cm.

"Landscape," vertical scroll ink painting by Chŏng Sŏn (1676–1759). In the University Museum, Seoul National University. 33 x 22 cm.

Korean art of the Chosŏn period



"Village Classroom," album leaf by Kim Hong-do, 18th century. Slight colour on paper. In the National Museum of Korea, Seoul. 28 x 24 cm.



Plate 14: (Top) Orion Press; (centre right, bottom right) National Museum of Korea, Seoul; (bottom left) Seoul National University, Korea; (centre left) Tenri Central Library, Tenri University, Nara, Japan



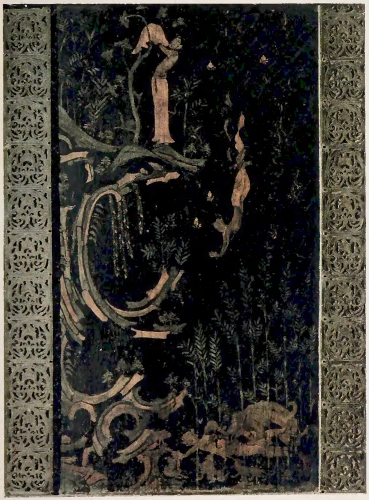
Gilt-wood Kuze Kannon, early 7th century, Asuka period. In the Yumedono ("Hall of Dreams"), Hōryū Temple, Ikaruga, Nara prefecture. Height 1.97 m.



The five-story pagoda of the Hōryū Temple, Ikaruga, Nara prefecture, wood and stucco, originally built in 607, reconstructed *c.* 680, Asuka and Hakuhō periods.

Japanese art of the Asuka and Hakuhō periods

Hungry Tigress jataka panel, detail from the Tamamushi Shrine, lacquer on wood with open metalwork borders, mid-7th century, Asuka period. In the Daihōzōden (Treasure Hall), Hōryū Temple, Ikaruga, Nara prefecture. Height of complete shrine 2.332 m.



The Yakushi Triad: Yakushi flanked by (left) Gakkō and (right) Nikkō, late 7th–early 8th century, Hakuhō period. Bronze with gilded mandorlas. In the Yakushi Temple, Nara. Heights of figures (left) 3.15 m, (centre) 2.55 m, (right) 3.18 m.



Plate 15: (Top left, bottom left) Shogakukan, Tokyo; (top right) Sybil Sassoon/Robert Harding Picture Library, London; (bottom right) Takichi Ine



"Kichijōten," painting on hemp cloth, 8th century, Nara period. In the Yakushi Temple, Nara. 53.3 x 32 cm.



Genji monogatari emaki ("Illustrated Tale of Genji"), detail of the hand scroll showing Prince Genji holding the infant Kaoru in section 3 of the "Kashiwagi" chapter of the novel by Lady Murasaki Shikibu, first half of the 12th century, Heian period. Ink and colour on paper. In the Tokugawa Art Museum, Nagoya.

Japanese art of the Nara and Heian periods

Painted clay Shukongōjin, 733, Nara period. In the Hokkedō (Sangatsudō), Tōdai Temple, Nara. Height 1.739 m.



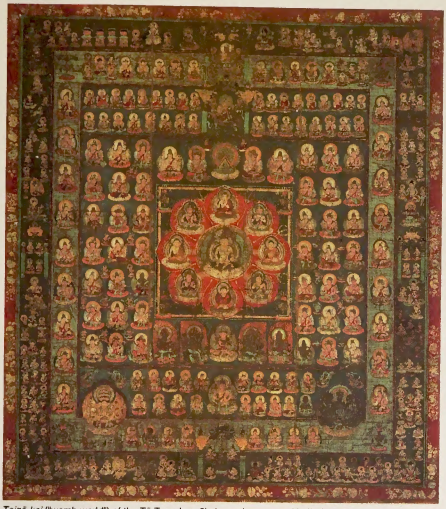
"The Death of Shaka," hanging scroll, colour on silk, 1086, Heian period. In the Reihokan (Treasure Hall), Koya-san, Wakayama prefecture. Height 2.69 m.



(left) Sakamoto Photo Laboratory, Tokyo; (top right) Tokugawa Reimeikai Foundation, Tokyo; (bottom ...ichiman Association of Mount Koya, Kongobu-ji, photograph, Shogakukan, Tokyo; (bottom left) ...to Laboratory, Tokyo

Amida Myorai by Jōchō, 1053, Heian period. Wood covered with gold leaf on a polychrome wood lotus pedestal. In the Hōō-dō ("Phoenix Hall") of the Byōdō Temple, Uji. Height 2.94 m.





Taizō-kai ("womb world") of the Tō Temple *ryōkai mandara*, second half of the 9th century, Heian period. Hanging scroll, colours on silk. In the Tō Temple (Kyōōgokoku Temple), Kyōto. 1.83 x 1.54 m.

Hōō-dō ("Phoenix Hall") of the Byōdō Temple, Uji, 1053, Heian period.



Plate 17: (Top left) Kyoogokoku-ji (To-ji), Kyoto, photograph, Benrido Co., Ltd., Tokyo; (top right, bottom) Sakamoto Photo Laboratory, Tokyo



Nandai-mon ("Great South Gate") of the Todai Temple, Nara. 1199, Kamakura period. Wood and stucco.



Amida (Amitābha), Buddha of the Western Paradise, wood, cut gold leaf, and polychromy sculpture by Kōshun and assistants, 1269, Kamakura period. In the Cleveland Museum of Art, Ohio, U.S. Height 94.6 cm.

"Landscape of the Four Seasons" (also called the "Longer Landscape Scroll"), detail of a hand scroll by Sesshū (1420–1506), Muromachi period. Ink and slight colour on paper. In the Mōri Museum, Yamaguchi, Japan. Height of scroll 40 cm.



Chinzō of Lan-ch'i Tao-lung (Japanese: Rankei Dōryū), hanging scroll by an unknown Japanese Zen monk, 1271, Kamakura period. Ink and light colour on silk. In the Kencho Temple, Kamakura. 1.05 m x 46.1 cm.



"The Burning of Sanjō Palace," detail from the *Heiji monogatari emaki* ("Hand Scroll of the Heiji War"), 13th century, Kamakura period. Ink and colour on paper. In the Museum of Fine Arts, Boston. Height of scroll 41.3 cm.



Plate 18: (Top left) International Society for Educational Information, Tokyo; (top right) © The Cleveland Museum of Art, John L. Severance Fund, 60.197; (centre) Bōfu Mōri Hoko-kai, Yamaguchi, Japan; (bottom right) the Museum of Fine Arts, Boston, Fenollosa-Weld Collection; (bottom left) Kencho-ji, Kamakura, Japan



Waterfall, detail of two panels of "Landscape of the Four Seasons," a pair of six-fold screens by Sesson Shūkei, second half of the 16th century, Muromachi period. Ink and light colour on paper. In the Art Institute of Chicago. Dimensions of one pair of screens 1.568 x 3.378 m.



"The Patriarch of Zen Buddhism," painting attributed to Kanō Motonobu (1476–1559), Muromachi period. Originally a door panel painting in the Daisen-in of Daitoku Temple, Kyōto, it is now mounted as a hanging scroll. Ink and colour on paper. In the Tokyo National Museum. 1.751 m x 88.4 cm.

Japanese art of the Kamakura and Muromachi periods

The priest Muchaku, painted wood sculpture by Unkei (1148–1223). Kamakura period. In the Kōfuku Temple, Nara. Height 1.893 m.



Ginkaku ("Silver Pavilion") of the Jishō Temple, Kyōto, 1489. Muromachi period. Wood and stucco with cypress bark roof.



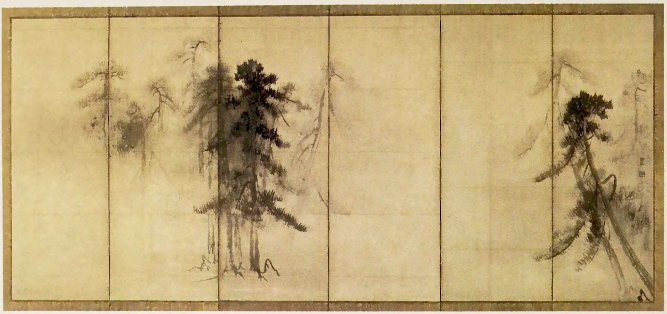
Plate 19: (Top left) The Art Institute of Chicago, gift of the Joseph and Helen Regenstein Foundation, 1958.168, photograph © 1994, The Art Institute of Chicago, all rights reserved; (top right) Collection of the Tokyo National Museum; (bottom right) © Dave Bartruff; (bottom left) Kofuku-ji, Nara, photograph, Asuka-en



Yomei-mon ("Gate of Sunlight") of the Tōshō-gū at Nikkō, dedicated to the shogun Tokugawa Ieyasu, 1636, early Tokugawa period. Carved, painted wood with gold leaf decoration.



Kosode (short-sleeved robe) of silk, decorated with a design of bamboo, 1573–1614, Momoyama period. In the Daihiko Senshu Bijutsu Kenkyūjo, Tokyo. Height 1.44 m.



"Pine Trees," one of a pair of six-panel screens by Hasegawa Tōhaku (1539–1610), Momoyama period. Ink on paper. In the Tokyo National Museum. Height 1.804 m.

"Cypress Trees," painting attributed to Kanō Kuninobu (Kanō Eitoku; 1543–90), Momoyama period. Formerly painted on sliding doors, it is now mounted as an eight-panel screen. In the Tokyo National Museum. 3.465 x 1.557 m.



Plate 20: (Top left) Orion Press; (top right) the Daihiko Senshu Bijutsu Kenkyujo, Tokyo; (centre) Collection of the Tokyo National Museum; (bottom) Benrido Co., Ltd., Tokyo



Himeji Castle, Hyōgo prefecture, built in the 14th century by the Akamatsu family, redesigned and rebuilt beginning in 1581 by the warlord Toyotomi Hideyoshi, and enlarged in 1601–09 by the Tokugawa family, Azuchi-Momoyama and Tokugawa periods.

Japanese art of the Azuchi-Momoyama and Tokugawa, or Edo, periods



"Southern Barbarians," one of a pair of six-fold screens attributed to Kanō Sanraku (1559–1635), Azuchi-Momoyama period. Colour and gold leaf on paper. In the Suntory Museum of Art, Tokyo. 1.82 x 3.71 m.



"Waves at Matsushima," one of a pair of six-panel screens by Sōtatsu, first half of the 17th century, Tokugawa period. Colour and gold leaf on paper. In the Freer Gallery of Art, Washington, D.C. 1.521 x 3.58 m.

Plate 21: (Top) Katsumi Tomita—Orion Press from SCALA/Art Resource, New York City; (centre) Suntory Museum of Art, Tokyo; (bottom) the Smithsonian Institution, Freer Gallery of Art, Washington, D.C.



"The Insistent Lover," woodblock print by Sugimura Jihei, 1680s. In the Art Institute of Chicago. 27.3 x 40.6 cm.



Enameled porcelain dish from Arita, Nabeshima ware, 17th century. In the Tokyo National Museum.

Japanese art of the Tokugawa period

"Yatsu-hashi Bridge and Irises," one of a pair of six-panel screens by Ogata Kōrin believed to date from 1701. Gold leaf and colour on paper. In the Metropolitan Museum of Art, New York City. 1.79 x 3.714 m.



Garden Kenroku, the garden of the estate of the Maeda family in Kanazawa, established in the 17th century and enlarged to its present size in the 19th century.



Plate 22: (Top left) The Art Institute of Chicago, Clarence Buckingham Collection, 1935.406; photograph © 1994; The Art Institute of Chicago, all rights reserved; (top right) Collection of the Tokyo National Museum; (centre) the Metropolitan Museum of Art, New York City, Louisa E. McBurney Gift Fund, 1963; (bottom) Hirai—Shostal/Superstock



"Verdant Hills and Scarlet Forests" from "Mist and Clouds," a set of album leaves by Uragami Gyokudō, *c.* 1811. Ink and colour on paper. In the Umezawa Memorial Hall, Tokyo. 28.5 x 21.9 cm.



"Thunder God" by Katsushika Hokusai, 1848. Ink and colour on paper. In the Freer Gallery of Art, Washington, D.C. 1.229 m x 49.5 cm.

"View from Komagata Temple near Azuma Bridge," woodblock print by Andō Hiroshige, 1857. In the Art Institute of Chicago. 36 x 24.1 cm.



"A Young Woman in a Summer Shower," woodblock print by Suzuki Harunobu, 1765. In the Art Institute of Chicago. 28.6 x 21.9 cm.



Plate 23: (Top left) Umezawa Memorial Hall, Tokyo; (top right) courtesy of the Freer Gallery of Art, Smithsonian Institution, Washington, D.C., 00.47; (bottom left) the Art Institute of Chicago, Clarence Buckingham Collection, 1925.3744, photograph © 1944, The Art Institute of Chicago, all rights reserved; (bottom right) the Art Institute of Chicago, Clarence Buckingham Collection, 1957.596, photograph © 1944, The Art Institute of Chicago, all rights reserved



"Professor Tenshin Okakura," hanging scroll by Shimomura Kanzan, *c.* 1922. Ink and colour on paper. In the Tokyo University of Arts. 1.36 m x 66.4 cm.



"Woman Combing Her Hair," woodblock print by Hashiguchi Goyō, 1920. In the Art Institute of Chicago. 44.8 x 34.9 cm.

Japanese art of the modern period



"Still Life of Salmon," oil on paper by Takahashi Yuichi, 1877. In the Tokyo University of Arts. 1.40 m x 46.5 cm.

"Reiko with a Woolen Shawl," oil on canvas by Kishida Ryūsei, 1921. In the Tokyo National Museum. 44.2 x 36.4 cm.



"Tzu-chin-ch'eng Palace," oil on canvas by Umehara Ryūzaburō, 1940. In the Eisei Bunko Foundation, Tokyo. 1.124 x 1.499 m.

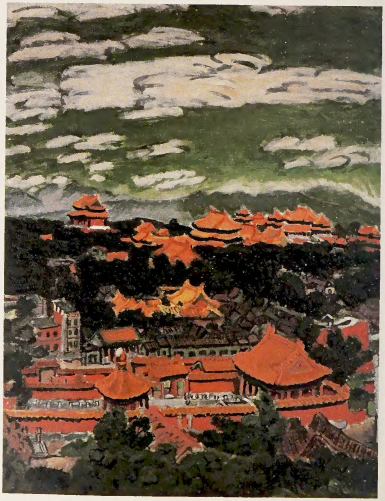


Plate 24: (Top left, top right) Tokyo University of Arts; (top centre) The Art Institute of Chicago, Clarence Buckingham Collection, 1928.646, photograph © 1994, The Art Institute of Chicago, all rights reserved; (bottom left) Collection of the Tokyo National Museum; (bottom right) Eisei Bunko Foundation, Tokyo



"A Group of Horses," detail of a four-fold painted screen by Kim Ki-ch'ang (1913–). Colour on paper. In the collection of the World House Gallery, New York. 2.73 × 3.64 m.
By courtesy of Won-Yong Kim

most commonly used, white porcelain alone was permitted as ritual ware for Confucian rites and ancestor worship.

Blue-and-white porcelain, inspired by early Ming models, appeared in Chosŏn Korea by the mid-15th century, and Chosŏn potters soon developed a distinct Korean or Chosŏn style of blue-and-white wares. Vessel forms are sturdy and simple; and the decoration, which is naive and refreshing, is kept to a minimum to emphasize the white background—a design tendency also observed on Chosŏn white porcelain with underglaze iron decoration. Chosŏn blue-and-white wares were produced by government-operated kilns in the Kwangju area near Seoul, mainly for palace and high government officials. In later years, vessels of low quality became accessible to commoners.

Modern period. The impact of modern Western art began to be felt during the last decades of the 19th century, when Korea was forced to enter into treaties with foreign governments. In 1900 a British architect, at the request of the Chosŏn government, designed the renaissance revival Tŏksu Palace in Seoul. The stone building, which later became the National Museum, was completed in 1909. With the construction of Western-style buildings in Seoul came the need for European furnishings. Glass was used in some doors in the palace and certain public buildings, and from 1900 electric lamps were installed. Just as the Koreans were beginning to familiarize themselves with such Western architectural concepts as spaciousness and convenience, the Japanese took over the government, bringing a new set of influences. During the 35-year period of the Japanese occupation (1910–45), some Koreans lived in strange, hybrid houses of Euro-Japanese style, but most clung to traditional Korean-style houses. Architects and carpenters capable of working in the traditional bracket system became so scarce that, after the liberation in 1945, the Korean government had to search out the few surviving older architects and set them to training younger ones, not so much to design new buildings in the traditional manner as to ensure that existing national cultural properties would be properly preserved.

European and Japanese influence

At the beginning of the Japanese occupation, traditional Korean painting was led by Cho Sŏk-chin and An Chung-shik. Cho was the last court painter of the Chosŏn dynasty, and An the last gentleman painter. But their styles were similar in their pursuit of the enervated southern Chinese style of the Ch'ing period, with its emphasis on fingertip technique. In 1911 the former Korean Imperial family set up an academy of painting to foster the traditional style, and, though it dissolved in 1919, a number of important painters were trained. By the 1930s the pattern of Korean painting began to change under the impact of both Japanese and European influences. In 1922 the Japanese had inaugurated an annual exhibition for Korean artists, designed to promote a new academic style. The only modern facilities for studying painting, whether Asian or Western, were Japanese. Despite the resistance of traditionalists, the Japanese impact was irresistible. Prominent painters during this period were Kim Ŭn-ho, Yi Sang-bŏm, Ko Hŭi-dong, Pyŏn Kwan-shik, and No Su-hyŏn. After World War II traditional painting began to assume a modern mode of expression, as may be seen in the works of a group of radical painters such as Kim Ki-ch'ang, Pak Nae-hyŏn, and Pak No-su. All of these artists were highly trained in the traditional mediums of ink and watercolour painting. Their paintings reflect a bold sense of composition and colour and also have the quality of genuine abstract art.

Introduction of oil painting

The introduction of the Western style via China in the 18th century had gone almost unnoticed. In 1899 the commissioning of a Dutch artist to paint the portraits of the king and the crown prince affronted the traditional court painters. When Ko Hŭi-dong returned from a period of study of oil painting in Japan, he was so ridiculed in public whenever he went out to sketch in oil that he finally gave up and returned to traditional painting. Nevertheless, several students followed his lead by going to Tokyo to study oil painting, and soon the new art became the dominant field of activity. Throughout the Japanese occupation, the main trend of Korean oil painting was the modest, representational school that had its roots in Impressionism. Among the outstanding painters in this style were Yi Chong-u, To Sang-bong, Kim Insŭng, and Pak Tŭk-sun. This academic-style painting filled the government-sponsored annual art exhibitions in Seoul into the early 1960s. But then the efforts of a group of nonrepresentational painters including Kim Hwan-ki, Yu Yŏng-guk, and Kwon Ok-yŏn reversed the trend and persuaded the government to establish an independent section in the exhibition for abstract art. The number of students studying abroad increased, and the influence of international-style painting progressed rapidly. In the last decades of the 20th century all kinds of European and American styles were introduced to artists in South Korea and experimented with. By contrast, in North Korea, artists have been restricted to traditional, conservative styles with which to represent an extremely limited range of subjects for the express purpose of promoting political propaganda. (W.-Y.K.)

Japanese visual arts

GENERAL CHARACTERISTICS

Most Japanese art bears the mark of extensive interaction with or reaction to outside forces. Buddhism, which originated in India and developed throughout Asia, was the most persistent vehicle of influence. It provided Japan with an already well-established iconography and also offered perspectives on the relationship between the visual arts and spiritual development. Notable influxes of Buddhism from Korea occurred in the 6th and 7th centuries. The Chinese T'ang international style was the focal point of

Japanese artistic development in the 8th century, while the iconographies of Chinese Esoteric Buddhism were highly influential from the 9th century. Major immigrations of Chinese Ch'an (Japanese: Zen) Buddhist monks in the 13th and 14th centuries and, to a lesser degree, in the 17th century placed indelible marks on Japanese visual culture. These periods of impact and assimilation brought not only religious iconography but also vast and largely undigested features of Chinese culture. Whole structures of cultural expression, ranging from a writing system to political structures, were presented to the Japanese.

Various theories have thus been posited which describe the development of Japanese culture and, in particular, visual culture as a cyclical pattern of assimilation, adaptation, and reaction. The reactive feature is sometimes used to describe periods in which Japanese art's most obviously unique and indigenous characteristics flourish. The notion of cyclical assimilation and then assertion of independence requires extensive nuancing, however. It should be recognized that, while there were periods in which either continental or indigenous art forms were dominant, usually the two forms coexisted.

Importance of nature

Another pervasive characteristic of Japanese art is an understanding of the natural world as a source of spiritual insight and an instructive mirror of human emotion. An indigenous religious sensibility that long preceded Buddhism perceived that a spiritual realm was manifest in nature. Rock outcroppings, waterfalls, and gnarled old trees were viewed as the abodes of spirits and were understood as their personification. This belief system endowed much of nature with numinous qualities. It nurtured, in turn, a sense of proximity to and intimacy with the world of spirit as well as a trust in nature's general benevolence. The cycle of the seasons was deeply instructive and revealed, for example, that immutability and transcendent perfection were not natural norms. Everything was understood as subject to a cycle of birth, fruition, death, and decay. (Imported Buddhist notions of transience were thus merged with the indigenous tendency to seek instruction from nature.)

Attentive proximity to nature developed and reinforced an aesthetic that generally avoided artifice. In the production of works of art, the natural qualities of constitutive materials were given special prominence and understood as integral to whatever total meaning a work professed. When, for example, Japanese Buddhist sculpture of the 9th century moved from the stucco or bronze T'ang models and turned for a time to natural, unpolychromed woods, already ancient iconographic forms were melded with a preexisting and multileveled respect for wood.

Union with the natural was also an element of Japanese architecture. Architecture seemed to conform to nature. The symmetry of Chinese-style temple plans gave way to asymmetrical layouts that followed the specific contours of hilly and mountainous topography. The borders existing between structures and the natural world were deliberately obscure. Elements such as long verandas and multiple sliding panels offered constant vistas on nature—although the nature was often carefully arranged and fabricated rather than wild and real.

Preference for imperfections

The perfectly formed work of art or architecture, unweathered and pristine, was ultimately considered distant, cold, and even grotesque. This sensibility was also apparent in tendencies of Japanese religious iconography. The ordered hierarchical sacred cosmology of the Buddhist world generally inherited from China bore the features of China's earthly imperial court system. While some of those features were retained in Japanese adaptation, there was also a concurrent and irrepressible trend toward creating easily approachable deities. This usually meant the elevation of ancillary deities such as Jizō Bosatsu or Kannon Bosatsu to levels of increased cult devotion. The inherent compassion of supreme deities was expressed through these figures and their iconography.

The interaction of the spiritual and natural world was also delightfully expressed in the many narrative scroll paintings produced in the medieval period. Stories of temple foundings and biographies of sainted founders were replete with episodes describing both heavenly and demonic forces roaming the earth and interacting with the populace on a human scale. There was a marked tendency toward the comfortable domestication of the supernatural. The sharp distinction between good and evil was gently reduced, and otherworldly beings took on characteristics of human ambiguity that granted them a level of approachability, prosaically flawing the perfect of either extreme.

Even more obviously decorative works such as the brightly polychromed overglaze enamels popular from the 17th century selected the preponderance of their surface imagery from the natural world. The repeated patterns found on surfaces of textiles, ceramics, and lacquerware are usually carefully worked abstractions of natural forms such as waves or pine needles. In many cases pattern, as a kind of hint or suggestion of molecular substructure, is preferred to carefully rendered realism.

The everyday world of human endeavour has been carefully observed by Japanese artists. For example, the human figure in a multiplicity of mundane poses was memorably recorded by the print artist Hokusai (1760–1849). The quirky and humorous seldom eluded the view of the many anonymous creators of medieval hand scrolls or 17th-century genre screen paintings. Blood and gore, whether in battle or criminal mayhem, were vigorously recorded as undeniable aspects of the human. Similarly, the sensual and erotic were rendered in delightful and uncensorious ways. The reverence and curiosity about the natural extended from botany to every dimension of human activity.

In summary, the range of Japanese visual art is extensive, and some elements seem truly antithetical. An illuminated sutra manuscript of the 12th century and a macabre scene of ritual disembowelment rendered by the 19th-century print artist Tsukioka Yoshitoshi can be forced into a common aesthetic only in the most artificial way. The viewer is thus advised to expect a startling range of diversity. Yet within that diverse body of expression certain characteristic elements seem to be recurrent: art that is aggressively assimilative; a profound respect for nature as a model; a decided preference for delight over dogmatic assertion in the description of phenomena; a tendency to give compassion and human scale to religious iconography; and an affection for materials as important vehicles of meaning.

STYLISTIC AND HISTORICAL DEVELOPMENT

Formative period. The terminology and chronology used in describing pre- and protohistoric Japan is generally agreed to be that of a Paleolithic, or Pre-Ceramic, stage dating from approximately 30,000 BC (although some posit an initial date as early as 200,000 BC); the Jōmon period (*c.* 10,500 BC–3rd century BC), variously subdivided; the Yayoi period (3rd century BC–3rd century AD); and the Tumulus, or Kofun, period (3rd century AD–AD 710).

Paleolithic stage. Until about 18,000 years ago, what is now known as the Japanese archipelago was connected to the East Asian landmass at several points. Similarly, the now divided islands were also joined at some points. These land passages account for the discovery of the remains of both prehistoric animals and microlithic cultures (but no pottery) of types usually associated with the continent. Continued warming trends, beginning about 20,000 years ago, eventually raised sea levels, thus cutting off all but the northern passage from Siberia.

The earliest human populations on the archipelago had subsisted on hunting and foraging, but with the warming trends the bounty of large, easy-to-fell animals began to die out while the variety and density of plant life rose dramatically. The increase in the number of sites discovered dating from 15,000 to 18,000 years ago suggests that once-roaming bands of hunter-gatherers were becoming gradually more sedentary and less dependent on foraging. As further evidence, the remains of charred cooking stones, indicating prolonged periods of use, have been discovered, and manufactured projectile points, including worked obsidian, dating from this period provide evidence of the people's adaptive skill in bringing down smaller, swifter game.

Mesolithic stage

Approximately 12,000 to 10,000 years ago, the definitive conditions for what is termed a Mesolithic stage became

apparent: a hunting culture employed microliths and, in addition, manufactured pottery. Just as the use of microlith weapons increased as a result of a decline in the numbers of big game, the manufacture of pottery was probably necessitated by a food supply crisis that required a means of storage and, perhaps, a method for boiling or otherwise cooking plants.

Jōmon period. Beginning in 1960, excavations of stratified layers in the Fukui Cave, Nagasaki prefecture in northwestern Kyushu, yielded shards of dirt-brown pottery with applied and incised or impressed decorative elements in linear relief and parallel ridges. The pottery was low-fired, and reassembled pieces are generally minimally decorated and have a small round-bottomed shape. Radiocarbon dating places the Fukui find to approximately 10,500 BC, and the Fukui shards are generally thought to mark the beginning of the Jōmon period. This early transitional period seems to lack convincing evidence of plant cultivation which would, along with microlith and pottery production, allow it to meet the criteria for a Neolithic culture.

The name Jōmon is a translation for "cord marks," the term the American zoologist Edward Sylvester Morse used in his book *Shell Mounds of Omori* (1879) to describe the distinctive decoration on the prehistoric pottery shards he found at Ōmori in southwestern Tokyo. Other names, such as "Ainu school pottery" and "shell mound pottery," were also applied to pottery from this period, but, after some decades, although cord marks are not the defining decorative scheme of the type, the term *jōmon* was generally accepted. The earliest stage of the period, to which the Fukui shards belong, has been given various names, including Incipient Jōmon and Subearliest Jōmon. Some scholars even call it Pre-Jōmon and argue that life during this stage showed only a slight advance from that of the Paleolithic. In 1937 Yamanouchi Sugao suggested the subdivisions Earliest, Early, Middle, Late, and Latest Jōmon for the remainder of the period. With refinements in chronology and the addition of some subsets, this terminology remains in use.

Earliest Jōmon

The period called Earliest, or Initial, Jōmon (*c.* 7500–5000 BC) produced bullet-shaped pots used for cooking or boiling food. The tapered bases of the pots were designed to stabilize the vessels in soft soil and ash at the centre of a fire pit. Decorative schemes included markings made by pressing shells and cords or by rolling a carved stick into the clay before it hardened. The shapes and worked surface textures of these early vessels suggest their probable precursors—leather, bark, or woven reed containers reinforced with clay. The Hanawadai site in Ibaraki prefecture constitutes the first recognized Earliest Jōmon community.

Early Jōmon (5000–3500 BC) sites suggest a pattern of increased stabilization of communities, the formation of small settlements, and the astute use of abundant natural resources. A general climatic warming trend encouraged habitation in the mountain areas of central Honshu as well as coastal areas. Remains of pit houses have been found arranged in horseshoe formations at various Early Jōmon sites. Each house consisted of a shallow pit with a tamped earthen floor and a grass roof designed so that rainwater runoff could be collected in storage jars.

Early Jōmon vessels generally continued the fundamental profile of a cone shape, narrow at the foot and gradually widening to the rim or mouth, but most had flat bottoms, a feature found only occasionally in the Earliest Jōmon period. The characteristic markings were impressed on damp clay with a twisted cord or cord-wrapped stick to produce a multiplicity of patterns. Other techniques, including shell impressions, were also used. In addition to the flared-mouth jars, shallow bowls and narrow-necked bottles were also introduced. The discovery of increasing varieties of flat-bottomed vessels appropriate for cooking, serving, and providing storage on flat earthen floors correlates with the evidence of the gradual formation of pit-house villages.

While pottery was the main form of visual expression in the Early Jōmon period, wood carving and lacquering are among the other significant forms of expression, suggesting the development of a more complex culture. Ropes, reed baskets, and wooden objects have been found at the Torihama mound site in Fukui prefecture. The oldest known examples of Japanese lacquerware—bowls and a comb—are also from this site.

Middle Jōmon

The Middle Jōmon period (3500–2500 BC) witnessed a dramatic increase both in population and in the number of settlements. Signs of incipient agriculture can be detected in this period, but this may have involved settling near wild plants and storing them effectively. Vessels began to take on heavy decorative schemes employing applied clay. The use of vessels for purposes beyond cooking and storage is also noted. Clay lamps, drum shells, and figurines strongly suggest an expanding use of the medium for religious symbolic expression. Fertility images of clay female figurines with exaggerated breasts and hips and of stone phalli have been located on stone platforms placed on the northwest side of dwellings. These platforms may represent early household altars. During this period jars were used for burial and were characteristically damaged so as to prohibit any other type of use.

Three distinct vessel styles were produced during the Middle Jōmon. The Katsusaka type, produced by mountain dwellers, has a burnt-reddish surface and is noted especially for extensive and flamboyant applied decorative schemes, some of which may have been related to a snake cult. The Otamadai type, produced by lowland peoples, was coloured dirt-brown with a mica additive and is somewhat more restrained in design. The Kasori E type has a salmon-orange surface. During this period a red ocher paint was introduced on some vessel surfaces, as was burnishing, perhaps in an attempt to reduce the porosity of the vessels.

In the Late Jōmon (2500–1000 BC) colder temperatures and increased rainfall forced migration from the central mountains to the eastern coastal areas of Honshu. There is evidence of even greater interest in ritual, probably because of the extensive decrease in population. From this time are found numerous ritual sites consisting of long stones laid out radially to form concentric circles. These stone circles, located at a distance from habitations, may have been related to burial or other ceremonies. Previously disparate tribes began to exhibit a greater cultural uniformity. Artifacts discovered in diverse coastal areas show a uniform vocabulary of expression and a consistent decorative system, suggesting more sophisticated methods of manufacturing, such as controlled firing of pottery, and increased specialization. The technique of erased cord marking, in which areas around applied cord marks were smoothed out, was increasingly used. This relates to a more general practice or interest in polished pottery surfaces. A unique black polished pottery type called Goryo has been found in central Kyushu. Some scholars suggest that this may in some way be imitative of Chinese black Lung-shan pottery (*c.* 2200–1700 BC).



Earliest Jōmon vessel, clay, *c.* 5000 BC. In the Hakodate Municipal Museum, Japan.
Height 16.6 cm.
Hakodate Municipal Museum, Japan

Latest Jōmon

Evidence from the Latest, or Final, Jōmon (*c.* 1000–3rd century BC) suggests that inhospitable forces, whether contagious disease or climate, were at work. There was a



Clay figurine dating from Latest Jōmon period, excavated at Ebisuda, Miyagi prefecture. In the collection of Tōhoku University, Sendai, Japan. Height 35.6 cm.
By courtesy of the Archaeological Seminar, Tohoku University, Sendai, Japan

considerable decrease in population and a regional fragmentation of cultural expression. Particularly noteworthy was the formation of quite distinct cultures in the north and south. The discovery of numerous small ritual implements, including pottery, suggests that the cultures developing in the north were rigidly structured and evinced considerable interest in ritual.

More than 50 percent of the Latest Jōmon sites are in northern Honshu, where significant quantities of polished or burnished pottery and lacquerware have been found. In fact, it is from this time that lacquer working—used for both decorative and waterproofing purposes—begins to emerge as a distinct craft. In general, the northern distinction between utilitarian and ritual ware became more pronounced, and the ritual ware became more elaborately conceived. The latter phenomenon is clearly illustrated by the unusual clay figurines with enormous goggle eyes that are characteristic of the Latest Jōmon.

In the south mobility and informality were the emerging characteristics of social organization and artistic expression. In distinction to the northern culture, the south seemed more affected by outside influences. Indeed, the incursions of continental culture would, in a few centuries, be based in the Kyushu area.

Yayoi period. In 1884 a shell mound site in the Yayoi district of Tokyo yielded pottery finds that were initially thought to be variants of Jōmon types but were later linked to similar discoveries in Kyushu and Honshu. Scholars gradually concluded that the pottery exhibited some continental influences but was the product of a distinct culture, which has been given the name Yayoi.

Both archaeological and written evidence point to increasing interaction between the mainland and the various polities on the Japanese archipelago at this time. Indeed, the chronology of the Yayoi period (3rd century BC–3rd century AD) roughly corresponds with the florescence of the aggressively internationalized Chinese Han dynasty (206 BC–AD 220).

The Yayoi culture thus marked a period of rapid differentiation from the preceding Jōmon culture. Jōmon, a hunting-and-gathering culture with possibly nascent forms of agriculture, experienced changes and transitions primarily in reaction to climatic and other natural stimulants. Yayoi, however, was greatly influenced by knowledge and techniques imported from China and Korea. The impact of continental cultures is decidedly clear in western Japan from about 400 BC, when primitive wet-rice cultivation techniques were introduced. Attendant to the emerging culture based on sedentary agriculture was the introduction of a significant architectural form, the raised thatched-roof granary. Bronze and iron implements and processes of metallurgy were also introduced and quickly assimilated, as the Yayoi people both copied and adapted types and styles already produced in China and Korea. Thus, while the decorative instincts of the Jōmon culture were limited primarily to the manipulation of clay, a variety of malleable materials, including bronze, iron, and glass, were increasingly available to artisans of the Yayoi period. The introduction of these various technologies, the development of a stable agricultural society, and the growth of a complex social hierarchy that characterized the period became the springboards for various forms of creative expression and provided increasing opportunities for the development of artistic forms.

Chinese and Korean influence



Earthenware jar excavated from Kugahara, Tokyo, Yayoi period. In the Tokyo National Museum. Height 36.3 cm.
By courtesy of the Tokyo National Museum

The Yayoi period is most often defined artistically by its dramatic shift in pottery style. The new type of pottery, reflecting continental styles, was made first in western Japan. It then moved eastward and became assimilated with existing Jōmon styles. Jōmon pottery was earthenware formed from readily available sedimentary clay and was generally stiff. Yayoi pottery was formed from a fine-grained clay of considerable plasticity found in the delta areas associated with rice cultivation. It was smooth, reddish orange in colour, thinly potted, symmetrical, and minimally decorated. The simpler, more reserved styles and forms emulated Chinese earthenware. It was also at this time that pottery began to be produced in sets, including pieces made for the storing, cooking, and serving of food.

In addition to the characteristic pottery that gave its site name to the period, the production of metal objects, particularly the *dōtaku* bells, represents a significant artistic manifestation of the Yayoi period. The *dōtaku* were cast in bronze and imitative of a Chinese musical instrument. More than 400 indigenously produced *dōtaku* have been discovered in Japan. These bells range from 4 to 50 inches in height. Their quality suggests a rather advanced state of technical acumen. Figural and decorative relief bands on these bells offer some, albeit highly interpretive, insights into Yayoi culture and suggest that shamanism was the dominant religious modality. The *dōtaku* appear not to have been used as musical instruments in Japan. Instead, like the bronze mirrors and other distinguished

Dōtaku bells



Dōtaku (bell-shaped bronze) with design of whirlpools, excavated from Uzumora, Hyōgo prefecture, Yayoi period. In the Tokyo National Museum.
By courtesy of the International Society for Educational Information, Tokyo

and precious implements transferred and adapted from Chinese and Korean forms, the *dōtaku* took on talismanic significance, and their possession implied social and religious power.

Tumulus, or Kofun, period. About AD 300 there appeared new and distinctive funerary customs whose most characteristic feature was chambered mound tombs. These tumuli, or *kofun* ("old mounds"), witnessed significant variations over the following 400 years but consistently dominated the period to which they gave their name. Some authorities have suggested that the development of these tombs was a natural evolution from a Yayoi-period custom of burial on high ground overlooking crop-producing fields. While partially convincing, this theory alone does not account for the sudden florescence of mound tombs, nor does it address the fact that some aspects of the tombs are clearly adaptations of a form preexisting on the Korean peninsula. Indeed, implements and artifacts discovered within these tombs suggest a strong link to peninsular culture.

Changes in tomb structure, as well as the quantity, quality, and type of grave goods discovered, offer considerable insight into the evolution of Japan's sociopolitical development from a group of interdependent agricultural communities to the unified state of the early 8th century. Of course, the material culture of the Kofun period extended far beyond the production of funerary art. For example, it is in this time that an essential form of Japanese expression, the Chinese writing system, made its appearance on the archipelago—a fact known from such evidence as inscribed metal implements. This system had a profound and comparatively quick influence not only on written language but also on the development of painting in Japan. Nevertheless, tombs are the repositories of the period's greatest visual achievements and are excellent indicators of more general cultural patterns at work. And, in that wider context, three distinct shifts in tomb style can be discerned that define the chronology of the period: Early Kofun of the 4th century, Middle Kofun covering the 5th and early 6th centuries, and Late Kofun, which lasted until the beginning of the 8th century and during which tomb burials were gradually replaced by Buddhist cremation ceremonies. The Late Kofun roughly coincides with the periods known to art historians as the Asuka (mid-6th century–645) and the Hakuhō (645–710).

Tombs of the Early Kofun period made use of and customized existing and compatible topography. When viewed from above, the tomb silhouette was either a rough circle or, more characteristically, an upper circle combined with a lower triangular form, suggesting the shape of an old-fashioned keyhole. The tombs contained a space for a wooden coffin and grave goods. This area was accessed through a vertical shaft near the top of the mound and was sealed off after burial was completed. The deceased were buried with materials that were either actual or symbolic indicators of social status. The grave goods were intended, as well, to sustain the spirit in its journey in the afterlife. They included bronze mirrors, items of jewelry made from jade and jasper, ceramic vessels, and iron weapons. Adorning the summit of the mound and at points on the circumference midway, at the base, and at the entrance to the tomb were variously articulated clay cylinder forms known as *haniwa* ("clay circle").

Haniwa sculpture

Haniwa were an unglazed, low-fired, reddish, porous earthenware made of the same material as a type of daily-use pottery called *haji* ware. These clay creations were shaped from coils or slabs and took the form of human figures, animals, and houses. The latter shape was usually set at the peak of the burial hillock. Many attempts have been made to interpret the function of *haniwa.* They seem to have served both as protective figures and as some type of support for the deceased in the afterlife. There is some suggestion that, similar to tomb figurines found in other cultures, they symbolized a retinue of living servants who might otherwise have been sacrificed upon the demise of their master. They are regionally distinctive and show a stylistic development from the decidedly schematic to realistic.



Haniwa horse, clay, 5th–6th century. In the Tokyo National Museum. Height 83.8 cm.
By courtesy of the International Society for Educational Information, Tokyo

Another type of ceramic prominent in the Kofun period was *sue* ware. Distinct from *haji* ware, it was high-fired and in its finished form had a gray cast. Occasionally, accidental ash glazing is found on the surface. Until the 7th century, *sue* ware was a product reserved for the elite, who used it both for daily ware and on ceremonial occasions. *Sue* ware was more closely identified with Korean ceramic technology and was the precursor for a variety of medieval Japanese ceramic types. Interestingly, both *haji* and *sue* ware found roles in funerary art.

After the 4th century, tomb builders abandoned naturally sympathetic topography and located mounds in clusters on flat land. There are differences in mound size, even within the clusters, suggesting levels of social status. The scale of these tombs, together with construction techniques, changed considerably. The tomb generally assumed to be that of the late 4th-century emperor Nintoku, located near the present-day city of Ōsaka, measures nearly 1,600 feet



Mausoleum of Nintoku, at Sakai, largest of the "keyhole" type tombs of the Tumulus, or Kofun, period. Circumference 2,718 m, height 21 m.
By courtesy of the International Society for Educational Information,Tokyo

in length and covers 80 acres (32 hectares). It is alternately surrounded by three moats and two greenbelts. Approximately 20,000 *haniwa* were thought to have been placed on the surface of this huge burial mound.

In the later part of the 5th century, the vertical shaft used to access the early pit tomb was replaced by the Korean-style horizontal corridor leading to a tomb chamber. This made multiple use of the tomb easier, and the notion of a family tomb came into existence. Also notable from the 5th century is the archaeological evidence of horse trappings and military hardware in tombs. *Haniwa* representing warriors and stylized military shields are also prominent. Records of diplomatic and military forays combine with the grave goods of the period to suggest a strong military cast to 5th- and 6th-century culture. However, in time these accoutrements of war and symbols of physical power are found in ancillary tombs rather than in the grave sites of known leaders. This suggests a gradual consolidation of power and the formation of a specialized military service within the kingdoms.

Japan's close relationship with Korean and Chinese cultures during the Kofun period effected an influx of peninsular craftsmen. This is reflected in the production of *sue* ware mentioned above and in the high quality of metalwork achieved. Mirrors are a particularly fine example of the development of metal craft. The typical East Asian mirror of the time is a metal disk brought to a high reflective finish on one side and elaborately decorated on the reverse. Such mirrors did not originate in Japan but seem to have been made and used there for religious and political purposes. The dominant Japanese creation myth describes the Sun Goddess, Amaterasu Omikami, being coaxed from hiding by seeing her reflection in a mirror. This may well have imparted a magico-religious quality to mirrors and caused them to be understood as authority symbols. Of particular note is the so-called *chokkomon* decorative scheme found on some of these mirrors and on other Early Kofun metalwork. *Chokkomon* means "patterns of straight line and arcs," and the motif has also been found chiseled on a wall in a Late Kofun tomb at the Idera tomb in Kyushu. It has been suggested that the abstract interweaving pattern may symbolize rope binding the dead to the tomb, an aspect of Chinese cosmology of the Han dynasty.

Metalwork

Late Kofun tombs are characterized by schemes of wall decoration within the burial chambers. Two especially important tombs have been excavated in the area just to the south of present-day Nara. The Takamatsu tomb (1972) and the Fujinoki tomb (1985) suggest high levels of artistic achievement and a sophisticated assimilation of continental culture. The Takamatsu tomb is noted for its wall paintings containing a design scheme representing a total Chinese cosmology. Included are especially fine female figure paintings. At Fujinoki exquisite and elaborate metalwork, including openwork gold crowns, a gilt bronze saddle bow, and gilt bronze shoes, was discovered. Design motifs show evidence of Chinese, Central Asian, and Indian sources.

Asuka period. The Asuka period was a time of transformation for Japanese society. It is named for the Asuka area at the southern end of the Nara (Yamato) Basin (a few miles to the south of the present-day city of Nara), which was the political and cultural centre of the country at the time. From there, the imperial court ruled over a loose confederation of rival clans, the most powerful of which were the Soga, Mononobe, and Nakatomi.

Japan's interest in and contacts with continental cultures continued to increase in the Asuka. A wide range of political and cultural relations with the Korean kingdoms of Koguryŏ, Silla, and, in particular, Paekche provided an opportunity for comparatively systematic assimilation of vast amounts of Korean culture, Chinese culture read through a Korean prism, and the religious beliefs of Buddhism. It was within that period of intensive relations with Paekche that critical foundations were constructed for a radical shift in the direction of Japanese visual arts.

Introduction of Buddhism

The most significant change, of course, was the introduction of Buddhism. Historians debate the actual date of the arrival of Buddhist texts, implements of worship, and iconography in Japan, but according to tradition a Paekche delegation to the emperor Kimmei in 538 or 552 made the presentation of certain religious articles. Given the extent of contact with Korea, however, various "unofficial" introductions of Buddhism had probably already occurred. The religion soon found favour in Japan and flourished under the powerful regent Prince Shōtoku (574–622), who established it as the state religion.

Buddhism was already a thousand years old when it arrived in Japan. It had transformed and been transformed by the iconography and artistic styles of the various cultures along its path of expansion from India. The central message of Gautama Buddha (6th–5th centuries BC) had also experienced multiple interpretations, as evidenced by the numerous sectarian divisions in Buddhism. The artistic forms necessary to provide the proper environment for the practice of the religion were well defined, however—calligraphy, painting, sculpture, liturgical implements, and temple architecture—and these were the means by which nearly all continental modes of Buddhism were absorbed and adapted by the Japanese culture.

During this period of intensive peninsular contact, Korean artisans skilled in metalwork, sculpture, painting, ceramics, and other fields necessary to the production of Buddhist iconography immigrated to or were brought to Japan in large numbers. While the practice of most of the above-mentioned forms was the purview of professionals, the calligraphic rendering of the written word was a skill available to the educated elite of the period. Thus, in the Asuka period the foundations of both individualized and public forms of visual expression were secured.

Architecture. Buddhism was established in Japan as a site-oriented faith. Temples with designs initially based on continental models became centres of worship. In contrast to the importance of funerary art in the Kofun period, the artistic expression of the Asuka period was developed within the matrix of public and privately commissioned temples. By the close of the Asuka period in the mid-7th century, nearly all vestiges of tomb burial customs were actually outlawed as the new faith made extensive inroads.

The most important temple complexes of the period are the Shitennō Temple at Ōsaka, the Wakakusa Temple near Nara (both constructed by Prince Shōtoku), and the Asuka Temple at Asuka. All three are known only through archaeological remains, although Wakakusa, Shōtoku's private temple, which was destroyed by fire in 670, was reincarnated as the Hōryū Temple (see below). These temple complexes replicated forms popular in Paekche and Koguryŏ. They were walled compounds in which stood a second rectangular compound bordered by a continuous roofed corridor. This second enclosure was entered through a central gate on its south side and contained a variety of internal structures, such as a pagoda (a form derived from the Indian stupa that served the dual functions of cosmological diagram and reliquary of important personages) and a Golden Hall (*kondō*), both used for worship. Support buildings, such as lecture halls, a belfry, and living quarters, lay outside and to the north of the inner cloister. True to the continental style, the buildings and gates were sited along a south-north axis and were symmetrical in layout. It was within the various buildings, particularly the *kondō*, that sculptures representing various figures in the Buddhist pantheon were placed.

Asuka temple layout

Roof tiles, stone, and cryptomeria wood were the essential building materials, all indigenous or locally produced. Structures relied on the placement of vertical wood pillars secured on finished stone bases. Horizontal elements were added in varying degrees of complexity, and structural balance was based on the essential pillar concept.

Sculpture. While the structures of these temples did not survive, certain important sculptures did, and these images are generally associated with the name of Kuratsukuri Tori (also known as Tori Busshi). Tori was descended from a family of saddlemakers. Excellence in this trade required mastery of the component media of lacquer, leather, wood, and metal, each of which was, in various ways, also used in the production of sculpture.

A large, seated, gilt-bronze image of Shaka (the Japanese name for Śākyamuni Buddha, the historical Buddha) survives from the Asuka Temple and is dated to 606. Also extant is the gilt-bronze Shaka Triad of Hōryū Temple, which is dated by inscription to 623. The Asuka Buddha, heavily restored, is attributed to Tori based on the stylistic similarity of its undisturbed head to the renderings found in the Shaka Triad, which is confidently assigned to the master sculptor's hand. A more controversial work is a gilt bronze Yakushi (the healing Buddha), which carries an inscription of 607. It is very close to the style of Tori, but many date the work to the latter part of the century. The Triad and the Yakushi are now housed in Hōryū Temple. An inscribed dedication found on the halo of the central figure of the Triad suggests that the ensemble was dedicated to the recently deceased Shōtoku and his consort. A stylistically related work is the wooden statue of the bodhisattva Kuze Kannon in the Yumedono ("Hall of Dreams") of the Hōryū Temple. The Tori style seen in these works reveals an interpretive dependence on Chinese Buddhist sculpture of the Northern Wei dynasty (386–534/535), such as that found in the cave sites at Lungmen. Symmetry, a highly stylized linear treatment of draped garments, and a reserved and gentle facial expression with a characteristic archaic smile are the prominent distinguishing features of this sculpture. The Japanese interpretations in bronze and wood advance the frontally focused Chinese relief sculptures by beginning to suggest more fully rounded figures.

Tori style



Shaka Triad by Tori, comprising the Buddha Śākyamuni and a pair of bodhisattvas, bronze, 623, Asuka period. In the *kondō* of the Hōryū Temple, Nara. Heights of figures: (left) 92 cm, (centre) 86.4 cm, (right) 93.9 cm.

Asuka-En

Painting. Buddhist temples were decorated not only with sculpture but also with religious paintings, tapestries, and other objects. Most such works from the Asuka period have not survived. An exception is the Tamamushi Shrine, which consists of a miniature *kondō* affixed to a rectangular pedestal or base. This assemblage of wood, metal, and lacquer provides an excellent view of what a *kondō* of the period may have looked like and, perhaps more important, is decorated with the only known painting from the Asuka period. The painting program on the miniature *kondō* seems to depict, through images on various panels and doors, the deities normally found in sculptural form within the hall. Paintings on the panels of the base show aspects of Buddhist cosmology and scenes from jataka tales, those narratives that tell of exemplary incidents in the previous incarnations of the Buddha. Perhaps best known is the jataka of the Hungry Tigress, in which the Buddha prior to enlightenment chances upon a tigress and her cubs starving in a desolate ravine and offers his own body to them. The painting depicts a sequential narrative in one panel, showing the saint removing his robe, leaping from a cliff, and being eaten by the tigers. The painting style suggests an Indian prototype vastly influenced by the fluid linearity of Chinese Wei styles.

Hakuhō period. In 645 Prince Nakono Ōe (later the emperor Tenji) and Nakatomi Kamatari (later Fujiwara Kamatari) led a successful coup and promulgated the Taika reforms, a series of edicts that significantly strengthened the control of the central government. Through successive regimes, some violently introduced, the structuring of a highly centralized government continued. A major feature of the centralization process was the incorporation

and use of Buddhism as an instrument of unification. The period was thus noted for a rapid expansion of Buddhism as aristocrats competed in the construction of temples. Increasing funds were allotted for the expansion of Buddhist temples and acquisition of the attendant iconography required for the expression of the faith.

The seat of government moved several times after the coup, but in 694 the court returned to the Asuka area and a plan to construct a permanent capital at Fujiwara was implemented. The capital was eventually moved again in 710 to Nara.

Art historians have given the name Hakuhō to the period beginning with the Taika reforms and ending with the imperial move to Nara. As noted, it overlaps with the Late Kofun period and is also sometimes referred to as the Late Asuka or Early Nara period.

Architecture. Four major temples, Asuka, Kawara, Kaikankai, and Yakushi, were already within the area of the planned capital site at Fujiwara. Of the four, only Yakushi Temple has survived, although not at Fujiwara but as an exact replica in Nara, constructed after the move of the capital in 710.

As an imperially commissioned temple, completed about 697, Yakushi had been very prominent at Fujiwara, and the relocated Yakushi Temple assumed equal importance when it was rebuilt at its new site (*c.* 730). Most recent evidence suggests that the Nara version of the temple was precisely faithful to the Fujiwara original and thus can be considered an example of late Hakuhō period temple design. Notable in its layout is the new prominence given to the *kondō* as a major structure; it is located in the centre of the compound flanked by two pagodas, which are afforded lesser importance than in earlier temple layouts. The *kondō* faced a large courtyard, and when its large central doors were opened, the assembled faithful were treated to an impressive view of the sacred images it housed. A unique feature of the Yakushi architecture is the use of the double-roof structure, in which a *mokoshi,* or roofed porch, was placed between two major stories.

Hakuhō temple layout

Despite Yakushi Temple's importance, Hōryū Temple, formerly Wakakusa, Prince Shōtoku's private temple, which was reconstructed about 680, remains the most significant extant repository of Asuka and Hakuhō art. By employing an asymmetrical layout, Hōryū differs dramatically from the axial-line layout of the major temples of the first half of the century. The gently tapering five-story pagoda and the wider, squatter *kondō* at Hōryū are placed adjacent to one another in the centre of the compound, their greatly varying sizes visually accommodated by an entry gate that is placed slightly off the central axis. This diversion from Chinese notions of balance became characteristic in many features of Japanese aesthetics.

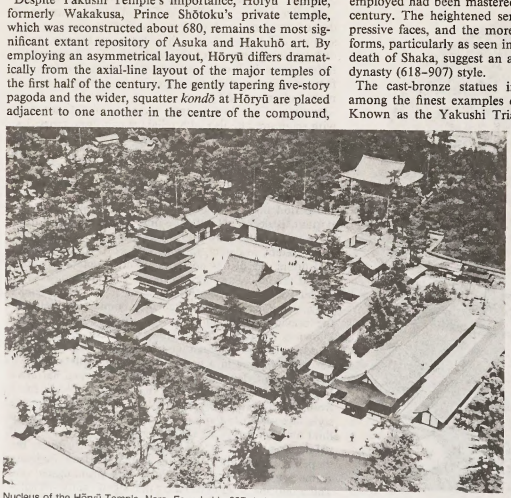
Sculpture. With the exception of the Shaka Triad dedicated in 623 (noted above), sculpture at Hōryū Temple was created in a period from approximately 650 until 711. Sculpture created from the middle of the century begins to reflect the influence of the Chinese Northern Ch'i dynasty (550–577) styles. The highly linear features of Northern Wei sculpture are supplanted by works that have emerged from their origin in relief wall sculpture and stand in the round as stolid, columnar figures with slight attenuation at the waist. Noteworthy of this new style are the four guardian figures who stand sentry over the quadrants surrounding the Shaka Triad and the more delicate Kudara Kannon held in the Hōryū Temple treasure house. The drapery at the feet of these statues flares forward rather than to the sides as in earlier works, allowing for a heightened sense of volume. The sculptures are executed in indigenous wood with some traces of gold and polychromy still remaining.

At Chūgū Temple, near Hōryū and once the residence of Prince Shōtoku's mother, a wood-sculpted image of Miroku Bosatsu embodies many of the characteristic features of the Hakuhō period. The delicately meditative figure sits with one leg pendant, its foot supported on a lotus, and the other leg crossed. The rounded cheeks, arching eyebrows, slight disproportionate swelling of the upper torso, and soft modeling suggest innocent, almost childlike features.

Other sculptural works from the second half of the 7th century show increased mastery of a wide variety of materials, including clay, and adaptive uses of lacquer. At Hōryū Temple a group of sculptures constructed of clay over wood and metal structures is arrayed in four distinct tableaux on the first level of the pagoda. Completed in 711, they are technically works falling into the Nara period. However, their virtuosity suggests that the techniques employed had been mastered in the final years of the 7th century. The heightened sense of realism, the more expressive faces, and the more rounded, three-dimensional forms, particularly as seen in the north-side tableau of the death of Shaka, suggest an assimilation of Chinese T'ang dynasty (618–907) style.

The Yakushi Triad

The cast-bronze statues in the Yakushi Temple are among the finest examples of Japanese sculpture extant. Known as the Yakushi Triad, the work consists of the



Nucleus of the Hōryū Temple, Nara. Founded in 607 during the Asuka period by Prince Shōtoku, the temple complex was destroyed by fire in 670 and rebuilt with a new layout in the Hakuhō period. In the courtyard are (upper left) the pagoda and (centre) *kondō* (Golden Hall), which houses the most sacred of the temple images.

By courtesy of the International Society for Educational Information, Tokyo

seated Yakushi Buddha flanked by the standing attendants Nikkō (bodhisattva of the Sun) and Gakkō (bodhisattva of the Moon). It is unclear whether these sculptures were produced after the temple's relocation to Nara or if they were transported from the original site. Literary evidence from the 11th century suggests the latter hypothesis, however, and these striking works are consistent with the confident, fleshy, idealized figures of the early T'ang period.

Painting. The finest examples of late 7th-century painting are found in the *kondō* at Hōryū Temple. Many of these wall paintings were irreparably damaged by fire in 1949, but photos and reproductions remain. One fresco depicting an Amida Triad shows graceful figures rendered with comparative naturalism and defined with consistent, unmodulated brush lines known as "wire lines" (*tessenbyō*). Like the Hōryū pagoda sculptures, the wall paintings suggest the influence of T'ang style.



Bodhisattva, detail from the Amida Triad, one of a series of frescoes in the *kondō* (Golden Hall) of Hōryū Temple, c. 710, Hakuhō period. Most of the original paintings were destroyed by fire in 1949. In the Hōryū Temple Museum, Nara. Height 3 m.
Horyu-ji Museum, Nara, Japan, photograph, Asuka-en

Nara period. During the reign of the empress Gemmei (707–715) the site of the capital was moved to the northwest sector of the Nara Basin. The new capital was called Heijō-kyō and is known today as Nara. Overcrowding, the relative isolation of the Fujiwara capital, and what would prove to be a constant nemesis to the Japanese state, an overly powerful Buddhist establishment, were some of the main factors contributing to the move.

The Nara period (710–784), also known as the Tempyō period, marks the apex of concentrated Japanese efforts to emulate Chinese cultural and political models. The new capital city was modeled after the T'ang capital at Ch'ang-an (near modern Sian), and complex legal codifications (*ritsuryō*) based on the Chinese system established an idealized order of social relationships and obligations. Thus, a hierarchical society was established, in symbolic and real terms, with all power proceeding from the emperor. The integration of religion into this scheme fixed a properly understood relationship between spiritual and earthly authority. Secular authority ultimately drew its power from this relationship.

The first several decades of the 8th century were marked by power struggles, political intrigue, attempted coups, and epidemics. This generally unsettled and contentious atmosphere caused the emperor Shōmu (724–749) to press determinedly for strengthening the spiritual corrective that he perceived to be offered by Buddhism. In 741 he established the *kokubunji* system, building a monastery and a nunnery in each province, all under a central authority at Nara. In 743 he initiated the planning for construction of that central authority—the Tōdai Temple—and of its central image, a massive bronze statue of the Birushana Buddha, known as the Great Buddha (Daibutsu). Shōmu envisioned religion as a supportive and integrated power in the rule of the state, not as a private faith or as a parallel or contending force. His merging of church and state, however, later enabled the temples to acquire wealth and privilege and allowed Buddhist priests to interfere in secular affairs, eventually leading to a degeneration of the national administration.

Integration of Buddhism and imperial rule

The Chinese taxation system, which was first adopted in Japan during the Taika reforms and further promulgated by the *ritsuryō* system, was based on the principle of state ownership of land and a national appropriation of the rice crop. It was, from the beginning, an inappropriate fit for the realities of Japanese agriculture. By mid-century the growth of privately owned, tax-free estates had shrunk the tax base, and this, coupled with the extraordinary demands for expansion, temple building, and icon manufacture, placed great strain on the general population. The concluding decades of the century were characterized by attempts to regularize government expenditure and to control the power of the Buddhist clergy. In 784 the capital was transferred north to Nagaoka, just west of present-day Kyōto. This was a prelude to the establishment of the capital at Heian-kyō, now called Kyōto, in 794.

What was meant to have been perceived as the cultural expression of a powerful government intent on adapting the very finest elements of T'ang international style was actually an extreme attempt by a comparatively weak government to conjure power through symbolic gesture. Nevertheless, the push to establish Japan as at least equal in splendour to T'ang China in its celebrations of Buddhism and to mark Japan as the magnificent easternmost extension of the faith's expansion in Asia allowed for a halcyon period for the creation of Buddhist art. Virtually all aspects of T'ang culture were absorbed during this period. Indeed, because Buddhism was later suppressed in China and much of T'ang Buddhist iconography destroyed, extant Japanese art of the Nara period serves as the single best reminder, once removed, of what the Buddhist glories of T'ang China must have been.

The main monument to the Nara period is undoubtedly the huge Tōdai Temple complex with its colossal central image of the cast-bronze Great Buddha. The construction of the Great Buddha Hall (Daibutsuden) commenced in 745, and dedication ceremonies for the nearly 15-metre-high seated figure were held in 752. Only fragments of the original are extant; most of the present sculpture dates to a reconstruction in 1692, which nevertheless gives ample sense of the scale and ambitions of Emperor Shōmu.

The Great Buddha

Two important Nara temples predate the initiation of the Tōdai Temple project. Kōfuku and Hokkedō were both constructed in the Gekyō ("Outer Capital") area to the east of the imperial palace (this "outer" area is now where most extant Nara period sites are located), and their assorted extant iconography bears witness to the revolution in sculptural rendering that is a distinguishing feature of 8th-century Japanese art.

Architecture. Kōfuku, the titular temple of the powerful Fujiwara clan, originally was established as Yamashina Temple in the area of present-day Kyōto in the mid-7th century. It was relocated to Nara in 710 by clan leader Fujiwara Fuhito (659–720) and given the name Kōfuku. In scale and in assembled iconography, Kōfuku Temple reflected the de facto political control wielded by the Fujiwara. Kōfuku was conceived as a place of worship and of monastic learning and as a centre for providing social services (such as medical and charitable aid) to the general population. After Fuhito's death an octagonal memorial hall was constructed, similar to the Yumedono at Hōryū Temple. This distinctive architectural addition to the tem-



Great Buddha Hall (Daibutsuden) of the Tōdai Temple, Nara. The original building was completed in 752 during the Nara period; the present hall is a late 17th-century reconstruction.
Orion Press—SCALA—Art Resource

ple indicated a shift away from the use of a pagoda or stupa as a large reliquary or memorial structure.

Sculpture. Records indicate that an assembly of 27 sculptures featuring images of the Shaka, bodhisattvas, and other attendants was completed and installed in Kōfuku Temple in 734. Of this grouping, six of an original ten disciples and all eight of the Eight Classes of Beings (designated as protectors or guardians of Buddhism) are extant.

Hollow-core dry-lacquer sculpture

These works are superb examples of the hollow-core dry-lacquer technique (*dakkatsu kanshitsu*) of sculpture, which was developed in China and enjoyed a sudden florescence in the Nara period. The technique required the creation of a rough clay-sculpted model on a wooden armature. This form was then covered with successive layers of lacquer-soaked hemp, each of which had to be dry before the next could be applied. Next, the back of the sculpture was cut open, the clay broken out, and, if necessary, a fresh armature inserted. Final surface refinements and details were then added using a paste mix of lacquer, sawdust, flour, and ground incense. Pigments and gold leaf were used to colour the finished form. Some sources suggest that the use of the new technique was encouraged in Japan because the casting of the Great Buddha at Tōdai Temple caused a shortage of the copper needed for bronze production. In addition, lacquer had the advantages of durability, insect resistance, and light weight. Perhaps most importantly, this additive technique of sculpting offered a more easily managed range of plastic expression.

The other major site for important Nara period works preceding the construction of Tōdai Temple is Hokkedō, also known as Sangatsudō, located at the eastern edge of the Tōdai complex. At present a curious mélange of 16 sculptural works is found on the altar platform in the temple. A hollow-core lacquer sculpture of the Fukūkenjaku Kannon functions as the central image. It is flanked by two clay images of the bodhisattvas Gakkō and Nikkō (sometimes identified as the guardian deities Bonten and Taishakuten). Much smaller than the central image, they date to the mid-8th century and were probably not created for the position that they now occupy. They are closely related stylistically to four clay guardian figures found in the ordination hall at Tōdai. Treatment of facial features in each of these clay works is individualized and highly refined. The Gakkō and Nikkō demonstrate a reserved energy and force while the guardian figures are bravura performances of gesture and elegant posture, but all are excellent examples of the Japanese command of T'ang-style powerful, inspirited, idealized forms.

The "secret" image of Shūkongōjin (733), a guardian deity, is secluded in a cordoned space behind the Fukūkenjaku Kannon and presented for viewing only once a year. A clay sculpture with its original gold leaf and polychromy largely intact, the thunderbolt-wielding deity is approximately life-size. Modeled on Chinese statues of guardian generals, the Shūkongōjin is a formidable image of swirling power and force and is the best preserved of the Nara-period clay sculptures, which like their hollow lacquer counterparts were formed on armatures.

Sculpture of the later Nara period began to employ yet another variation of the lacquer technique, that of adding lacquered cloth over a carved wood core (*mokushin kanshitsu*). Paste techniques similar to those used for hollow-core lacquer sculpture enhanced the image, and some elements were occasionally constructed solely of lacquer disguised as wood. To alleviate splitting caused by expansion and contraction, the wood core was usually partially hollowed. The use of lacquered wood-core techniques may reflect an attempt to reduce the expense involved in previously described sculptural methods. It also indicated an increasing penchant for employing wood, an abundant natural resource.

Wood-core lacquer sculpture

The new technique may have been brought to Japan by Chinese artists accompanying the venerable Chinese monk Ganjin (Chinese: Chien-chen) (688–763), for whom Tōshōdai Temple (founded in 759) was constructed, by some accounts from a structure disassembled and moved from the imperial palace. Housed in Tōshōdai Temple are several works using the new wood-core lacquer technique, including a 534-centimetre-high, 11-headed, 1,000-armed Senju Kannon, as well as a hollow-core dry-lacquer sculpture of Ganjin and a Birushana Buddha of the same medium, both dating to about 760. The Ganjin sculpture is a particularly commanding work that embodies the authority and dignity of the aged, blind patriarch.

In addition to new construction techniques, sculpture of the late Nara period also shows a stylistic shift, probably imitating a continental trend, toward more mannered depictions of drapery and a more stolid, fleshy form, conveying a brooding feeling. Typical is the rendering of a tight-fitting garment at the thighs of a subject, with drapery elsewhere carved in evenly spaced, concentric waves. This style, *hompa-shiki,* came to greater prominence in the early Heian period.

Painting. Painting of the period emulated T'ang prototypes. Noteworthy is an image of the deity Kichijōten, housed in Yakushi Temple. This work on hemp depicts in full polychromy a full-cheeked beauty in the high T'ang style, which was characterized by slightly elongated, pleasantly rounded figures rendered with long curvilinear brushstrokes. A horizontal narrative scroll painting, *Kako genzai inga kyō* ("Sutra of Cause and Effect"), depicts in crisp primary pigments and a naive, almost childlike style

events in the life of the historical Shaka Buddha as well as various incidents in his previous incarnations. This work features painting on the upper register and explanatory text beneath. It stands at the head of a particularly fruitful tradition in Japanese painting types.

Decorative arts. Located within the Tōdai complex, to the northwest of the Great Buddha Hall, is the Shōsō-in treasure house, an imperial storage house constructed shortly after the death of Emperor Shōmu in 756. The joined-log structure, built of cypress timbers that are triangular in cross section, resembles a granary, a style of construction known as *azekura-zukuri.* It houses an accumulation of imperial objects as well as gifts received at the dedication of the Great Buddha and later donated by Emperor Shōmu's consort, Empress Kōmyō. Additional articles were added to the collection in the middle of the Heian period (794–1185). The core group donated by Empress Kōmyō totaled about 600 objects, including calligraphy, paintings, religious ritual implements, samples of medicines, mirrors, lacquerware, and masks. The objects received as gifts at the dedication of the Great Buddha have origins as distant as the Mediterranean basin. Most of the objects seem to be of Japanese origin, but they reflect a range of T'ang period styles, and they provide a vivid picture of T'ang and Nara decorative arts.

Shōsō-in treasure house

One of the few decorative art forms not well represented in the Shōsō-in treasure house is ceramics. Nara period ceramics, like the other arts, were imitations and adaptations of T'ang styles. Of note was the production of wares covered with a lead glaze of the T'ang *san ts'ai,* or three-colour, type (green, brown, and yellow), a two-colour type (green and white), and a monochrome green.

Heian period. In 784 the emperor Kammu (737–806) relocated the seat of government to Nagaoka, a site to the north of Nara and slightly to the west of present-day Kyōto. This move was an attempt to escape the meddling dominance of the Buddhist clerics in Nara and thus to allow unfettered development of a centralized government. Nagaoka was marred by contention and assassination, however, rendering it an inauspicious location for the capital. Thus, in 794 a site to the east of Nagaoka on a plain sheltered on the west, north, and east by mountains and intersected by ample north-south rivers was judged appropriate by geomancers. Named Heian-kyō ("Capital of Peace and Tranquility") and later known as Kyōto, this city was modeled on the grid pattern of the T'ang Chinese capital at Ch'ang-an. Heian-kyō remained the site of the imperial residence, if not the consistent seat of political power, until 1868.

For nearly four centuries Heian-kyō was the crucible for a remarkable florescence of Japanese art. Within a century after the move from Nara, political chaos in China caused the cessation of official embassies to the continent. Free from the overwhelming dominance of Chinese artistic models, Japanese culture, particularly literature and the visual arts, was able to evolve along independent lines and reflect national concerns. These developments were invigorated through dedicated aristocratic patronage of both religious art and a nascent secular art.

Florescence of Japanese art

Although sometimes viewed nostalgically as an unbroken series of halcyon years during which courtly aestheticism produced the "classical" body of Japanese literature and art, the Heian period was, in fact, a time of on-going political contention during which imperial attempts at centralization of government were consistently checked and ultimately defeated by powerful provincial warlords. In theory, all land and its revenue-producing capability was the property of the central government. In reality, outlying land managers, aristocrats, temples, and warlords accumulated landholdings unabated throughout the Heian period, ultimately crippling the economic power of the court. In the waning years of the 12th century, internal strife over succession and a scramble for what wealth remained in imperial hands forced the court to restore order with the assistance of the warrior class. This steady decline in aristocratic fortune and power was perceived by courtiers as an impending collapse of a natural and just order.

Literature and art of the period were thus often infused with nuances of sadness, melancholy, and regret. The consolations of Buddhism stressed the impermanence of life and served to reinforce for aristocratic believers the deeper meaning of readily apparent social developments. Indeed the shifting emphases found in Buddhist iconography during the Heian period are incomprehensible unless viewed in the context of doctrinal responses to social change. Most significant among these are the establishment of two Japanese schools of Esoteric Buddhism, Tendai and Shingon, in the early 9th century, the increasing appeal of Amidism in the 10th century, and, with the understanding that Buddhism entered a final millenarian era in the mid-11th century, a florescence of various iconography produced in the hopes of gaining religious merit.

Esoteric Buddhism. The court in Heian-kyō was justifiably wary of Buddhism, at least in any powerfully institutionalized form. Thus, in the configuration of the new capital, only two Buddhist temples were allowed within the boundaries of the city. Tō Temple and Sai Temple, located respectively at the east and west side of Rashomon, the southern gateway to Heian-kyō, were conceded space that was as far away as possible from the imperial palace and government offices in the north of the capital.

Dissatisfaction with the scholastic Buddhism of the Nara sects was also voiced by some clerics. An imperially approved embassy to China in 804 included the well-known monk Saichō (767–822) and the lesser-known Kūkai (774–835). Saichō studied the teachings of the T'ien-t'ai sect (Japanese: Tendai), which emphasized the impermanence of all things, an ultimate reality beyond conceptualization, and a fundamental unity of things. Meditational practices were believed to lead to enlightenment. The *Lotus Sutra* (Japanese: *Myōhō renge kyō*) was regarded as the primary text of the sect. Saichō returned to Japan in 805 and petitioned the court to establish a Tendai monastery on Mount Hiei overlooking Heian-kyō. His request was granted, but the emperor required Saichō to include some Esoteric practices in his Tendai system.

Saichō and Kūkai

Kūkai devoted himself to the mastery of the relatively new Esoteric Buddhist beliefs under the Chinese master Hui-kuo. Returning to Japan in 806, more than a year after Saichō (who regarded Esoteric teachings as an aspect of the more inclusive Tendai tradition), Kūkai was welcomed as an Esoteric master. Through the force of his personality and the attraction of his teachings, he eclipsed Saichō in popularity. Whatever particular differences are found between Tendai and Shingon, as Kūkai's syncretic doctrine is called, the two schools are grouped under the central category of *mikkyō,* or Esoteric Buddhism. Neither belief system, as interpreted in Japan, rigorously emulated the Chinese versions; they were syntheses created by Saichō and Kūkai.

Esoteric Buddhism relied heavily on visualization in its praxis. The creation of an environment of worship was essential. The use of mandalas, expressed both in two dimensions as paintings and in three dimensions as ensembles of sculpture, invited the believer into a diagrammatic rendering of a spiritual cosmos. A central tenet of Esoteric teaching was the nonduality of the Buddha. Whatever the manifestations, the phenomenal and the transcendental are the same. The goal of spiritual practice was to unite what seemed to the uninitiated to be separate realms. Thus, one of the most important iconographic images was the *ryōkai mandara* ("mandala of the two worlds"), which consisted of two parts—the *kongō-kai* ("diamond world") and the *taizō-kai* ("womb world")—that organized the Buddhist divinities and their relationships in a prescribed gridlike configuration. The deities or spiritual entities portrayed in these paired paintings represent, in the *kongō-kai,* the realm of transcendant, clear enlightenment and, in the *taizō-kai,* the humane, compassionate aspects of the Buddha. It was the repetitive meditative practice of journey through and visceral assimilation of this symbolic, schematic cosmos that could lead the believer to an enlightenment of unity.

Use of mandalas

In 823 Kūkai was granted imperial permission to take over the leadership of Tō Temple (also known as Kyōōgokoku Temple), at Heian-kyō's southern entrance. Images developed under his instruction probably included forerunners of the particular *ryōkai mandara* known as the

Tō Temple mandala. Stylistically, these paintings reveal a shift from T'ang painting style to a flatter, more decorative approach to image. Also in the sanctuary at Tō Temple is an important assemblage of sculpture that constitutes a three-dimensional mandala. In a tandem similar to the one effected in mandala painting, dual aspects of the single Buddha nature are portrayed. Bodhisattvas represent limitless compassion, while other assemblages portray yet another dimension of the central divinity, one that came to heightened prominence in Shingon practice, the fierce Myō-ō, or Kings of Bright Wisdom. These manifestations, perhaps best typified by Fudō Myō-ō, are terrifying and uncompromising guides for the believer in the journey to enlightenment. To the unfamiliar eye, their appearance seems demonic, but their wrath is directed at the enemies of Buddhism. They extend to a more fantastic perceptual level the role of guardian general deities and offer a realistic assessment of the intensity of dedication needed for enlightenment.

In general, sculpture produced in the 9th and 10th centuries followed and developed from the techniques of the late Nara period. Many works were constructed using variations of the lacquered wood-core technique. The heightened mannerism and heavy, brooding quality noticeable in some late Nara works are found in abundance in the early Heian period. The great late 8th-century standing Yakushi figure housed at Jingo Temple north of Kyōto perhaps best typifies this style. Other fine examples can be found in Murō Temple, a well-known Esoteric sanctuary to the east of Nara. Stylistically, these works hearken back to a type of sandalwood sculpture that enjoyed popularity in India and in China. With occasional elaborations through the use of lacquer, these powerful works were essentially carved from large, single pieces of wood, a technique called *ichiboku-zukuri*. It has been suggested that Buddhist reformers planned the contrast between the abrupt, extreme force of these sculptures and the aristocratic elegance of Nara period works. Created unabashedly of wood, they represented the elemental force of the forests that surrounded the urban centres.

Single-block wood sculpture

Because Esoteric practitioners were initially relegated to the mountainous regions outside the capital, the layouts and architecture of their temples varied greatly from the flatland architecture of the Nara temples and, thus, from the symmetrical Chinese styles. Placement and structure were adapted to rugged terrain, creating unique solutions. Ironically, this relative individualism of style was a subtle symbolic disruption of Nara period attempts at a hierarchically dispersed power through visual means.

The highly syncretic nature of Esoteric Buddhism considered the noumenal aspects of indigenous religions as emanations or manifestations of the Buddha essence. Rather than confronting and competing with native deities and belief systems, *mikkyō* readily adapted and included their features. For example, Shintō, the primary indigenous religion, which had developed from ancient animistic cults, had a very limited iconographic program. Until the Heian period, Shintō deities (*kami*) were largely considered to be unseen, often formless spirits that inhabited or personified such natural phenomena as the sky, mountains, and waterfalls. Esoteric Buddhism, however, encouraged the inclusion of Shintō deities in a kind of subordinate tandem with Buddhist deities in a variety of visual representations. This incorporation of Shintō *kami* not only served as an acknowledgment of indigenous beliefs but also increased the thematic scope of Buddhist art, particularly landscape painting. The Shintō belief that topography and its included features of rivers, trees, and distinctive rock formations were the abodes of the spirits meant that a sacred formation of mountains could be interpreted as a topographic mandala. Rendering these forms in painting expanded the iconographer's repertoire beyond the production of anthropomorphized deities. This theory in which *kami* are viewed as temporary manifestations of the essential Buddha, allowing each Shintō deity to be identified with a Buddhist one, is known as *honji-suijaku*.

Honji-suijaku

Amidism. Devotion to Amida Buddha, who presided over the Western Paradise, or Pure Land, began in Japan within the *mikkyō* sects, and in the 10th century Amida worship began to gain momentum as a distinct form of Japanese Buddhist belief.

Like Esoteric Buddhism, Amidism encouraged an iconography that formed a total ambience of worship. The focus of faith in Amida was rebirth in the Western Paradise. Therefore, painted and sculpted representations of that celestial realm were produced as objects of consolation. Paintings from the Nara period of the Amida and his Western Paradise are geometrically ordered descriptions of a hierarchical world in which Amida is enthroned as a ruler. In mid-Heian Amidist images, the once-ancillary image of the descending Amida takes on central prominence. This image of the Amida Buddha and attendants descending from the heavens to greet the soul of the dying believer is called a *raigōzu* ("image of coming to greet"). The theme would later be developed during the Kamakura period as an immensely popular icon, but it saw its first powerful expressions during the Heian period in the late 11th century. As is typical of Amidism, the compassionate attitude of the divinity superceded expressions of awesome might. Amidism differed significantly in emphasis from Esoteric Buddhism in that it did not require a guided initiation into mysteries. An expression of faith in the Amida Buddha through the invocation of his name in the *nembutsu* prayer was the single requirement for salvation. Iconography served mainly as a reminder of the coming consolations rather than as the tool for a meditative journey to enlightenment.

Raigō images

One of the most elegant monuments to Amidist faith is the Phoenix Hall (Hōō-dō) at the Byōdō Temple in Uji, located on the Uji River to the southeast of Kyōto. Originally used as a villa by the Fujiwara family, this summer retreat was converted to a temple by Fujiwara Yorimichi (990–1074) in 1053. The architecture of the building, including the style and configuration of its interior iconography, was intended to suggest a massive expression of *raigō* imagery, whether viewed by a worshiper within the sanctuary or by a visitor approaching the complex from a distance. Viewed frontally, the hall resembles a large bird with its wings extended as if in landing, recalling the downward flight of the Amida and bodhisattvas who welcome the faithful. Contained in the breast of this great creature is the sanctuary, where a magnificent Amida sculpture by Jōchō (d. 1057), the premier sculptor of the period, rests on a central altar. Positioned on the surrounding walls is an array of smaller wood-sculpted *apsaras* (heavenly nymphs) playing musical instruments and riding on stylized clouds. Traces of poorly preserved polychrome painting on the interior walls depict not only the expected *raigō* scene but also the gently rolling topography of central Japan, suggesting that the court-sponsored painting bureau had developed a strong indigenous expression which now supplanted Chinese models in religious iconography.

The Jōchō Amida sculpture, one of the most sublime expressions of Amidist belief, marks the ascendancy of a new style and technique in sculpting. Serene, unadorned, reserved yet powerfully comforting, this image is composed of numerous wood pieces that have been carved and hollowed, then joined together and surfaced with lacquered cloth and gold leaf. This joined-block construction technique (*yosegi-zukuri*) allowed for a sculpture lighter in feeling and in fact, but it generally precluded the deep and dramatic carving found in single-block construction. Thus, the exaggerated, mannered presentations of Esoteric sculpture of the previous centuries were supplanted by a noble, evenly proportioned figure, and scale and calm mien replaced drama as a means to engage the believer.

Joined-block wood sculpture

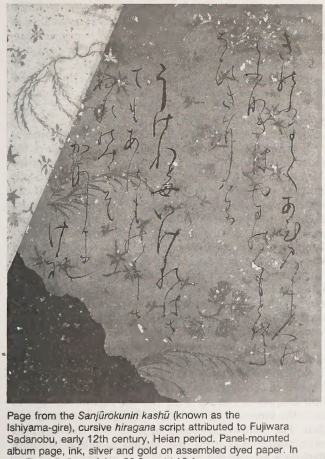
In 985 the Tendai monk Genshin (942–1017) produced the 10-part treatise *Ōjō Yōshū* ("Essentials of Salvation"), a major synthesis of Buddhist theory on the issues of suffering and reward and a pragmatic guide for believers who sought rebirth in the Western Paradise. Genshin described in compelling detail the cosmology of the six realms of existence of the Impure Land (*rokudō*) in an effort to encourage people to strive to achieve rebirth in the Pure Land of Amida. Genshin's descriptions of hell and its tortures were particularly influential as a source for artists in meeting a demand for graphic images of hell intended for meditation and instruction of the faithful.

Calligraphy and painting. The break in regular communication with China from the mid-9th century commenced a long period of fruitful development in Japanese literature and its expression through the mediums of calligraphy and painting. Calligraphy of the Nara period was known for its transmission and assimilation of the major Chinese writing styles, as well as for some forays into individualized expression and adaptation of technical features of character representation. Modified versions of Chinese characters, known as *man'yōgana,* were employed to represent Japanese phonetic sounds, and two even more abbreviated phonetic writing systems, *hiragana* and *katakana,* were known in nascent form. The former was highly stylized and cursive, while the latter was somewhat more severe and rectilinear in form. Use of *hiragana* was relegated to women, while men continued to control the learning and use of the traditional Chinese characters. However, during the Heian period *hiragana* was recognized as an official writing method, and an integrated use of the adapted Chinese characters (*kanji*) and *hiragana* became a widely accepted form of written expression.

The Buddhist monk Kūkai was an important calligraphic stylist and was posthumously recognized as the patron of calligraphers. His highly expressive and mannered presentation of characters was seen and admired in official correspondence, but, more significantly, he employed the brush in a spiritual exercise of rendering important sutra texts or single, meaning-laden *kanji*. These explorations functioned as part of an Esoteric rite that approximated use of a personalized mandala. Kūkai forcefully established the link between word and image embodied in a calligraphy text, and his work served as an important catalyst in the Heian period, when the rendering of a *kanji* or a phonetic symbol came to be appreciated not as an illustrative gesture but as a form of expression multivalent in its epistemological potential. In ensuing decades and centuries courtiers expanded on his work and explored the potentials suggested not just in a single character but in whole, secular texts, mainly poetry.

Importance of poetry

Contemporary documents discuss the relationship between poetry and painting. Poems were used as the subject of paintings, and calligraphers often wrote poems on paintings or on specially prepared square papers (*shikishi*) later affixed to a painting. Although virtually no examples of this custom survive from the Heian period, it is known through documentary sources and through revivals of the practice in subsequent centuries. Poetry was also inscribed on elaborately decorated sheets of paper which were preserved as individual units, consolidated in albums, or arranged on horizontal scrolls. The early 12th-century *Sanjūrokunin kashū* ("Anthologies of Thirty-Six Poets") is perhaps the finest Heian example of verse executed on sumptuously prepared and illustrated papers. The preeminence of the calligraphic word in interpretive union with painting or as a thematic inspiration for painting was a hallmark of the Heian period.



Page from the *Sanjūrokunin kashū* (known as the Ishiyama-gire), cursive *hiragana* script attributed to Fujiwara Sadanobu, early 12th century, Heian period. Panel-mounted album page, ink, silver and gold on assembled dyed paper. In the Freer Gallery of Art, 20.3 cm × 16.1 cm.

The Smithsonian Institution, Freer Gallery of Art, Washington, D.C.

Changes in painting technique evident in the Heian period may well have been the result of the general and rapidly growing development of sophisticated calligraphic skills. The T'ang Chinese method of employing the even iron-wire brush line to delineate forms was gradually supplanted in the 11th century by subtle introductions of modulated, calligraphic brushwork, engendering greater liveliness in form, particularly in the renderings of such grand subjects as *raigō* and the Buddha's entry into nirvana.

Important secular works from the 11th century, such as *Shōtoku taishi eden* ("Illustrated Biography of Prince Shōtoku") and the Senzui folding screens (*byōbu*), also reveal the development of indigenous painting styles within the original interpretive matrix of Chinese forms. Although the Chinese method of representing narrative in a landscape setting is honoured, with each narrative episode shown in a discrete topographic pocket, the topography and other telling elements take on the appearance of Japanese rather than Chinese surroundings. By the end of the Heian period, a clear distinction could be made between paintings using Chinese themes and styles and those with Japanese subjects and techniques, with the former known as Kara-e and the latter as Yamato-e.

Kara-e and Yamato-e

Some of the most celebrated examples of Yamato-e are the horizontal narrative hand scrolls (*emaki* or *emakimono*) produced in the 12th century. This format, which had been introduced from China in the 6th or 7th century, had already been used effectively in Japan, most notably for the Nara period *Kako genzai inga kyō* ("Sutra of Cause and Effect"), but these early scrolls are thought to be imitative of Chinese works. In the Late Heian, however, *emaki* began to develop a unique Japanese character and proved to be particularly well suited to Japanese expression.

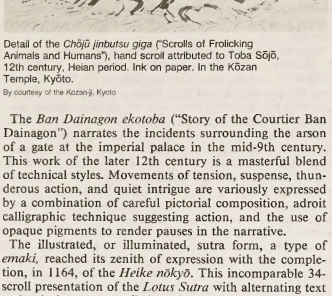
There are few extant narrative scrolls dating from the Heian period. Their quality is extraordinary, however, and probably representative of a larger number of works no longer extant. Typically, the format of presentation was that of alternating bodies of text and painting. The best of these works were not ploddingly literal in their visual interpretations of text. Rather they were carefully selective of their points of illustration, allowing maximum freedom to the viewer's imagination and demonstrating a complementary rather than repetitive use of text and image.

The range of expressive technique available to artists was considerable, and adaptation of style and composition to suit the tone of a narrative was, judging from available evidence, astute. *Genji monogatari emaki,* an illustrated narration of the late 10th- or early 11th-century court romance *Genji monogatari* (*The Tale of Genji*), was produced in the first half of the 12th century. The tale, which relates the life and loves of Prince Genji, is undergirded with Buddhist metaphysics and is thought to offer an approximate fictional description of court life at the time of its composition. It provides a complex analysis of emotions that are always obliquely expressed because of the constraints of court etiquette. The artists thus convey mood not with facial expression or gesticulation, which would violate the highly refined court aesthetic, but with formally posed figures rendered in opaque pigments and the skillful use of depicted architectural elements. Treatments of interior space subtly suggest the emotion masked by the human figures.

Genji monogatari emaki

Quite different from *The Tale of Genji* scroll is the 12th-century *Shigisan engi emaki* ("Legends of Shigisan Temple"). Drawing on folkloric sources, it is a tale of miracles attributed to the Shingon monk Myōren, who

resided on Mount Shigi near Nara in the latter part of the 9th century. The uninhibited depiction of action and movement central to various episodes is rendered by lively and varied brush strokes. Similarly, the first scrolls of the *Chōjū jinbutsu giga* ("Scrolls of Frolicking Animals and Humans"), products of the 12th century (later scrolls are dated to the 13th century), satirize human foibles through the depiction of anthropomorphized animals rendered in masterfully vibrant ink monochrome brushwork.



Detail of the *Chōjū jinbutsu giga* ("Scrolls of Frolicking Animals and Humans"), hand scroll attributed to Toba Sōjō, 12th century, Heian period. Ink on paper. In the Kōzan Temple, Kyōto.
By courtesy of the Kozanji, Kyoto

The *Ban Dainagon ekotoba* ("Story of the Courtier Ban Dainagon") narrates the incidents surrounding the arson of a gate at the imperial palace in the mid-9th century. This work of the later 12th century is a masterful blend of technical styles. Movements of tension, suspense, thunderous action, and quiet intrigue are variously expressed by a combination of careful pictorial composition, adroit calligraphic technique suggesting action, and the use of opaque pigments to render pauses in the narrative.

The illustrated, or illuminated, sutra form, a type of *emaki,* reached its zenith of expression with the completion, in 1164, of the *Heike nōkyō*. This incomparable 34-scroll presentation of the *Lotus Sutra* with alternating text and painting was an offering of the military leader Taira Kiyomori (1118–1181).

Decorative arts. The kilns at Sanage to the east of present-day Nagoya provided functional ceramic pieces for the court. These were largely forms and glazes that were imitative of Chinese three-colour and celadon potteries, which used lead in their glazes. Lacquerware emerged as an art that provided a means of producing the effect of inlay work popular mainly as an import item during the Nara period.

The Kamakura period. From the middle of the 12th century the reality of true imperial court control over Japan was largely a fiction. The Taira (Heike), a provincial warrior family, assumed the role of imperial protector and became the effectual power wielder. From that time they entered into a protracted struggle for hegemony with the Minamoto (Genji), a powerful clan from eastern Japan. The Gempei War between the families raged through much of Japan's central island from 1180 to 1185, during which such major temples as Tōdai and Kōfuku and their contents were completely destroyed. The Minamoto eventually emerged victorious, and, under the leadership of Minamoto Yoritomo, the culture and structure of national leadership shifted from the civil aristocracy to the hands of a provincial warrior class. In 1192 Yoritomo was named *seii taishōgun* ("barbarian-quelling generalissimo") by the court, thus initiating an office of military dictator that would persist until the Meiji Restoration in 1868. Yoritomo located his power centre (later termed shogunate, or *bakufu,* literally "tent government") in Kamakura, a small seaside village on a peninsula to the south of present-day Tokyo. Control of the shogunate soon passed to the Hōjō family through Yoritomo's widow, but the government did not return to Kyōto until 1333. The years from 1185 to 1333 are thus known as the Kamakura period.

Rise of the warrior class

The military victory and subsequent structural changes not only established the new ruling group in a position of military and economic power but also allowed for the infusion and development of a new cultural ethos—one that paralleled but was determinedly distinct from that developed by the court in Nara and in Kyōto. Warrior values of strength, discipline, austerity, and immediacy found resonance in the practices of Zen Buddhism. This strain of Buddhism had long played a subsidiary role in Japan, but, from the 13th century, strong Japanese adherents were bolstered in number and authoritative leadership by immigrant Chinese monks who had been displaced by the Mongol conquests in China. Zen Buddhism offered the new military leadership a nonthreatening alternative to the Tendai-controlled religious establishment that dominated the Kyōto court. The iconographic needs and the inherent aesthetic predispositions of Zen Buddhism were refined through this initial relationship with the Kamakura elite and over the next several centuries became widely influential throughout Japan.

Populist religious movements, particularly those generated by Amidist beliefs during the Heian period, grew even stronger and more diverse during the Kamakura period, increasing demands for Buddhist iconography. During the 13th century fears of an invasion by the Mongols from the mainland were realized on two occasions (1274 and 1281). Both times the invaders were repulsed, but these episodes and their anticipation contributed to a pervasive anxiety that was more than occasionally exhibited in the mood and theme of religious iconography. It was a time punctuated by prayers of supplication and pleas for divine intervention. Although quite different in their fundamental precepts, the simple and direct means of access to salvation or enlightenment offered by Zen or Amidist practices were exceedingly popular.

Sculpture. Perhaps no single feature of the Kamakura period so exemplifies the unique character of the age as does the emergence of bold new sculptural styles. As a result of the widespread destruction wrought by the Gempei War, it was necessary to replace the extensive loss of religious sculpture. The most compelling works of the period were created in the 13th century, notably by the Kei family, led by Kōkei and his son Unkei. Inspired both by the exquisite idealism of the Nara-period works and by the fashion for realism found in Chinese Sung dynasty sculpture, the best of Kamakura-period sculpture conveyed intense corporeal presence. The style is frequently referred to as "Kamakura realism" but should not be confused with the notion of "realistic" in the sense of faithful rendering of the natural. While, for example, there is reference to careful anatomic understanding, this understanding is often rendered in extreme statement. The huge guardian figures created by Unkei and other Kei artists to flank the Nandai-mon ("Great South Gate") at Tōdai Temple are the epitome of this style. With bulging eyes, limbs lined with tributaries of protruding veins, and theatrical poses, these and similar works were direct and accessible to the mass of the Buddhist faithful.

"Kamakura realism"

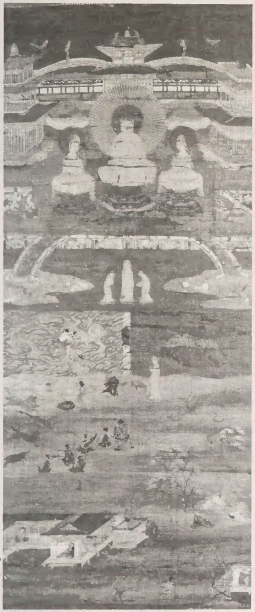
In portraying a range of divine concerns from protection to sympathetic consolation, Kamakura sculpture responded to the spiritual climate of the age. The sculpture by Unkei's son Kōshō (d. 1237) of Kūya, the rugged old mendicant who advocated the unceasing repetition of the *nembutsu* prayer, is depicted realistically as determined and gnarly but with the fantastic grace note of a string of small Amida figures emerging from his mouth—a literal representation of his teaching. An exquisitely refined evocation of the protective and welcoming presence of the Amida is seen in the sculpture dated to 1269 and a product of the atelier of Kōshun. With its surface completely adorned with gold-leaf pattern cuttings (*kirikane*), this figure proclaims celestial splendour. The intensity of the deity's gaze, omniscient and direct, is accomplished by a Kamakura-period innovation: inlaid crystal eyes backed by white paper appropriately coloured to effect iris and pupil. For Kōfuku Temple, Unkei sustained the remarkable standards of the temple's renowned Nara-period hollow-lacquer sculpture with his production of figures such as the famous disciples of the Buddha, Muchaku and Seshin. The portrait sculpture of Muchaku conveys firm resolve, seasoned realism, and, thanks to subtle handling of fleshiness around the eyes, a hint of humour.

The finest of Kamakura-period sculpture is a seamless

union of meticulously crafted and assembled parts. While wood was the medium of choice, the dominant presence of a single tree, a feature characteristic of early Heian sculpture, is no longer present. The joined-block method was used with much greater frequency than in previous periods. Effects were achieved through the coordination of skills, and specialization within workshops was common. In some cases the face of a sculpture was worked separately, as if a mask, and then affixed to the sculpture. The refinement of this ability to work on individual parts allowed for remarkable detail and expressive effects, enhancing the meticulous realism characteristic of Kamakura sculpture.

New architectural styles

Architecture. New architectural styles also emerged from the void created by the Gempei War devastation. No person was more instrumental in the renaissance of religious art and architecture than the monk Shunjōbō Chōgen (1121–1206), who oversaw the restoration of Tōdai Temple. Nandai-mon, the main entry gate of this revered temple, offers a superb example of the *tenjiku-yō* ("Indian style," although it originated in Southern Sung China) of architecture introduced during the reconstruction. Extravagantly conceived eaves wing out more than 5 metres (16 feet), supported by nine-tier brackets. The simple mechanics of this operation lie exposed, straightforward and bold, like the overall impression of the gate's design. Far less refined than Heian architecture, the immediacy of the new, and comparatively short-lived, style typified the aesthetic directness of the age.



"White Path to the Western Paradise Across Two Rivers" hanging scroll, *c.* 1300, Kamakura period. Ink, gold and silver, and kirikane on silk. In the Cleveland Museum of Art, Ohio, U.S. 123.4 cm × 50.7 cm.

© The Cleveland Museum of Art, gift of the Norweb Foundation, 1955

Similar simple lines were features of the newly introduced Chinese Ch'an (Japanese: Zen) religious architectural style, which included slightly more complex bracketing supports joining columns and horizontal elements. Prosaic elements such as dormitories and refectories were part of the central plan, thus uniting overall design scheme with the important realities of communal life. Meditation halls were also more prominent. On the whole, however, traditional architecture in the period tended toward the decorative and overworked, as nonessential elements multiplied and functional units were embellished. Oddly, where technical virtuosity served the sculptural format well, the necessary simplicity of monumental architecture was compromised by too much display.

Painting. Painting of the Kamakura period, both religious and secular, was marked by a sense of immediacy and vitality. The Amidist sects spawned cults that emphasized devotion to particular intercessory figures who had initially been considered ancillary in the overall Pure Land Buddhist pantheon. The popularity of Amidism also encouraged the creation of elaborately conceived spiritual cosmologies in paintings depicting the six realms of existence. In a variation of that theme, paintings of the Nika Byakudō ("White Path to the Western Paradise Across Two Rivers") type show both the difficulties encountered by the believer journeying to the Western Paradise and, at the centre, the bodhisattva Jizō benevolently ministering to those in need. Similarly, *raigō* paintings featuring depictions of the Amida and entourage descending from paradise to greet the souls of the recently deceased faithful enjoyed considerable popularity.

As was the case with sculptural representation, immediacy and accessibility were the most desired attributes of religious iconography. Religious foundations made extensive use of the narrative scroll format to honour sect anniversaries or histories and to document the biographies of founders and other major personalities. In vitality of defining brushwork, rich palette, and lavish depiction of the sundry details of contemporaneous existence, such works serve as essential records of the material culture of the Kamakura period; but in a more profound religious sense, they are visual evidence of the strong Japanese penchant for grounding the spiritual experience in the easily approachable guise of everyday life.

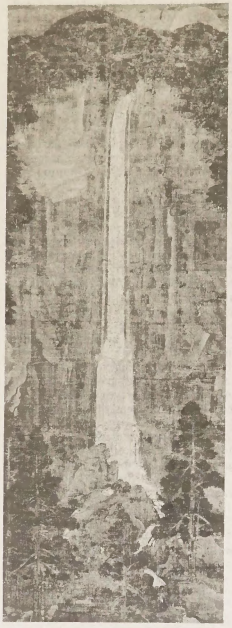
Zen Buddhist painting

The use of iconography in Zen Buddhism was not as extensive as in other sects, but mentor and patriarch portraiture played a significant role in the ritual of the transmission of teaching authority. Here, too, the penetrating effect of presence was the quality most sought in these visages. Ink monochrome painting was also employed by Zen adepts as a form of participatory spiritual exercise. In addition to representations of personages or historic moments, real or legendary, associated with Zen, Zen painters also depicted subjects not obviously religious in theme. Bird-and-flower paintings were created and queried for insights into spiritual meaning, and gradually the landscape painting offered accretions of symbolic meaning indicative of internal, spiritual journeys.

During the Kamakura period, Buddhism continued and strengthened systematic efforts to incorporate the indigenous religion, Shintō, by identifying local gods and numinous presences as manifestations of Buddhist deities. This system was called *honchi-suijaku,* and its principles were applied extensively. Religious paintings often depicted the figures of both Buddhist and Shintō manifestation in some mandala-like format. Likewise, Buddhist paintings, especially of the *honchi-suijaku* type, frequently incorporated Shintō sacred sites into their landscapes. Not precisely of this type, but a sublime derivative, is the icon of Nachi Falls. Here, a sacred site on the Kii Peninsula south of Ise reveals the haunting presence of the great, constantly plunging force which all but overwhelms the small architecture of the Shintō shrine that honours the natural site. Thus, certain Buddhist traditional painting techniques revealed the sacredness of adopted territory.

Secular painting

In the realm of secular painting, as in the religious world noted above, the narrative scroll continued to develop as an essential expressive format. The popularity of war tales, appropriate to the climate inspired by the interests of



"Nachi Falls," Shintō hanging scroll painting, 13th–14th century, Kamakura period. Colour on silk. In the Nezu Art Museum, Tokyo.
By courtesy of the Nezu Art Museum, Tokyo

the new national leadership and by the threat and reality of foreign invasion, is readily apparent in extant paintings commemorating various domestic martial episodes. Few paintings of the period capture the force, confusion, and terror of battle as effectively as does the episode of the burning of the Sanjō Palace in the *Heiji monogatari emaki*. Here, the artist uses highly animated, modulated strokes of defining ink, judicious, repetitive patterning, and the application of opaque colour to produce a series of carefully joined vignettes that intimately and actively tell the story.

The court, although stripped of political power, continued to be an arbiter of cultural matters. Most especially, it dominated the development of a national literature and the rendering of that literature in relation to painting and calligraphy. The various modes of joining word and image continued to be the specialized purview of aristocratic culture. In the early 13th century important anthologies were assembled of the works of the 36 ancient poets who had been "canonized" in the Heian period, and portraits of these masters were popular painting subjects. Often, the horizontal narrative scroll format was used to present the poets as if they were engaged in poetry competitions, composing linked verse (*renga*), with representative verse juxtaposed by their images. Thus, even the comparatively subdued ambience of court culture was animated by the format so attuned to the dynamism of the period. The 36-poet genre was thereafter a resilient theme and a standard way of expressing high literary reference in painting.

Secular portraiture saw developments stimulated in part by the central role of patriarch and mentor portraits in the Zen tradition. The schematic or generalized visages of the Heian-period indigenous traditions were influenced by these imported developments. Court and military portraits of the period tend to present the subject in the stiff, opaque, and decorative surrounding typical of Heian style, but faces are more realistically and individually rendered.

The Muromachi period. Ashikaga Takauji, a warrior commissioned by the Kamakura shogun to put down an attempt at imperial restoration in Kyōto, astutely surveyed circumstances and, during the years 1333 to 1336, transformed his role from that of insurrection queller to usurper of shogunal power. The Muromachi period (1338–1573) takes its name from a district in Kyōto where the new shogunal line of the Ashikaga family established its residence. With Takauji's ascendancy a split occurred in the imperial lineage. A southern court in exile formed in the Yoshino Mountains, to the south of Nara, while a court in residence, under the Ashikaga hand, ruled from Kyōto. This double regency continued until the end of the century, when a duplicitous compromise finally stripped the southern court of claims to power. This imperfectly resolved situation henceforth provided both political and romantic aesthetic evocations of legitimate power deposed. It became a rallying point for royalists and a continuing subtle undercurrent in literature and the visual arts, a metaphor for the contention between the brute force of arriviste pretensions and the sublime culture of legitimate rule. By extension, it harked back to the halcyon days of Heian court rule.

The Ashikaga family held relative control of national power until the mid-15th century, when other aggressive provincial warlords provoked a struggle that culminated in the Ōnin War (1467–77). This civil war laid ruin to much of Kyōto and was, in effect, the initial skirmish in a century of ongoing military conflict. Ashikaga men continued as figurehead rulers until 1573, when Oda Nobunaga dismissed the last Ashikaga shogun.

Military patronage of the arts

The military rulers attempted to establish their legitimacy through their patronage of the arts. They assiduously promoted Zen Buddhism and Chinese culture in opposition to the aristocratic preference for indigenous styles. The increase in trade with Ming China and the avid cultivation of things Chinese encouraged by the Ashikaga rulers established a dominant aesthetic mode for the period, and journeys of monk-artists to and from China provided yet another avenue for stimulation of the arts.

Meanwhile, Japanese court culture, using Heian-period aesthetic achievements as a canonical norm, continued to foster and develop indigenous visual forms. Both court and shogunal currents—what might be called, respectively, conservative and Sinophilic—were strengthened by interaction. While the various patronage groups were, to a degree, antagonistic, the juxtaposition generally stimulated experiment and challenged stagnant modes of visual representation.

In addition to the cultural changes wrought by sheer military power, the egalitarian structures of Zen Buddhism and other populist Buddhist movements provided the possibility of startlingly swift advancement and important patronage for talented but low-born individuals. Many found that the indeterminate social status afforded by religious ordination provided the means to move freely among different classes. It was also common to assume a religious status as a kind of social camouflage without the actual benefit of ordination. Artists of every sort found temple ateliers congenial to their talents in this time of relative meritocracy.

Zen Buddhism firmly established its role of intellectual leadership during the Muromachi period and provided a strong line of continuity with the aesthetic trends established during the Kamakura period. Growing in real power, the temples became to an even greater degree centres for the consideration, assimilation, and dissemination of continental culture.

Buddhism responded to the elevated cultural aspiration of its believers, clerics and laity alike, by providing occasions in which the realms of the aesthetic and religious were, in practice, joined. The development of the tea

ceremony, which became increasingly important because it linked heightened religious sensibility with artistic connoisseurship, is a prime example of Buddhism's role in fostering new art forms in this period.

Regional dissemination of central cultural values was another important catalyst for development. The increasing strength of provincial leaders allowed them to assume patronage roles and to invite distinguished Kyōto artists to regions distant from the centre of culture. From the time of the Ōnin War and in the century following, this process was accelerated as Kyōto was engulfed in martial conflict.

Painting and calligraphy. The most significant developments in Japanese painting during the Muromachi years involved the assimilation of the Chinese ink monochrome tradition, known in Japanese as *suiboku-ga* or *sumi-e*. Zen Buddhism was the principal conduit for knowledge of this painting tradition, which was originally understood as an exercise potentially leading to enlightenment, either through viewing or in the practice of putting brush to paper. It was practiced both by amateurs and by professional monk-painters in temple ateliers.

Chinese ink monochrome painting



"Catching a Catfish with a Gourd," Zen hanging scroll painting by Josetsu, with laudatory inscription (not reproduced here) by Daigaku Shuso and other priests, *c.* 1413, Muromachi period. Ink and faint colour on paper (*suiboku-ga*). In the Taizō-in, Kyōto.
By courtesy of the Taizo-in, Kyoto, Japan

In about the year 1413 Josetsu, a monk-artist of the Ashikaga-supported Shōkoku Temple, was commissioned by Ashikaga Yoshimochi (1386–1428) to produce a painting in the "new style" (thought to be that of the Southern Sung). The resulting work shows a man with a gourd standing near a stream and a catfish swimming in the water. Originally mounted as a small screen, the painting was soon transferred to the hanging scroll format, and the poetic commentaries of 30 monks were appended to the painting. This is perhaps the most famous work by the artist, who—as the master of Shūbun (fl. 15th century), who, in turn, instructed Sesshū (1420–1506)—is generally considered to stand at the head of the most important lineage of Muromachi ink painters. Josetsu's work alludes to the shogun's dominance of the elemental and sometimes unpredictable forces of nature and society, which are represented by the wily catfish. It can be understood as an instruction in the limitations of and deluded aspirations for power. It also suggests a style of Zen pedagogy in which a visual or verbal puzzle (in this case, how does one catch a slippery catfish with a small gourd?) prompts a dialogue between master and pupil as an exercise toward enlightenment. Noteworthy here is the fact of an exceptionally skilled painter operating well within the parameters of painting as religious exercise and also revealing the essential links between political power and Zen Buddhism's florescence.

Later, ink monochrome painters attempted themes that included Taoist and Buddhist patriarchal and mythical subjects, bird-and-flower compositions, and landscapes. It is instructive to note that in the course of the 15th century the progress of the three-generation lineage of Josetsu, Shūbun, and Sesshū can be described as a movement from physical permanence and relative security to a peripatetic existence necessitated by political instability and from conservative to more generalized or secular themes. Sesshū, who traveled to Ming China and was influenced by court painters, saw that Chinese painting was far greater in range than the ink monochrome tradition. His later works demonstrate a subtle use of colour and complex, seemingly random compositional formats, suggesting an increasing priority of brush stroke and patterning as the true subject. His long landscape scroll produced for the Mori clan in Yamaguchi is a brilliant study of boldly described forms in linear movement. He is also known for his landscapes in the *haboku* ("splashed-ink") technique, a style promulgated by Chinese Ch'an Buddhist painters who likened the spontaneous brushwork and intuitively understood (rather than realistically depicted) forms to the spontaneous, intuitive experience of Ch'an enlightenment.

Ink painting was not only the province of Zen Buddhists. Painters of the Ami lineage (so called because they used the suffix "ami" in their names to indicate their faith in Amida) served the Ashikaga shoguns as aesthetic advisers.

Painters of the Ami lineage



"Viewing a Waterfall," hanging scroll by Geiami, 1480, Muromachi period. Ink and colour on paper. In the Nezu Art Museum, Tokyo. 106 cm × 30.3 cm.
Nezu Art Museum, Tokyo

They graded and organized the shogunal collections of Chinese art and, as practitioners of the ink monochrome form, tended to a more gentle, polished conservatism than the bold, rough brushwork of the Shōkoku Temple painters. This tendency is seen in a work by Geiami (1431–85) painted on the occasion (*c.* 1480) of the departure of his pupil Kenkō Shōkei. It depicts the common subject of travelers passing beyond a turbulent pool and plunging waterfall to a temporary shelter nestled in a grotto. The sentiment is clear, and the execution reveals a mannered, controlled hand. The standard representation of receding far distance is suggested, but, in comparison with Chinese and earlier Japanese works, the balance of the painting is now subtly disrupted and the frontal plane becomes the focus of the work.

Sesson Shūkei (1504–*c.* 1589), another master of the ink monochrome format, was a mendicant with eclectic training. Eschewing any apprenticeship in Kyōto, he worked in a style charged with highly individualistic energy that captured the brooding uncertainties of the warring period.

The late Muromachi transition to secularization of the ink monochrome format is best expressed in the work of Kanō Motonobu (1476–1559). His father, Masanobu (1434–1530), stands at the head of a lineage that became, in following centuries, the dominant Japanese painting academy. The Kanō group was one of several important ateliers to develop important syntheses of Chinese and indigenous painting styles. Motonobu married into the Tosa family of Yamato-e painters, symbolically and literally effecting this gradual eclecticism. His sliding door panel paintings for Daitoku Temple in Kyōto depict famous episodes of Zen enlightenment. High professionalism, delicate coloration, and a skillful narrative instinct are apparent in this sweeping composition.

Although ink monochrome painting reached its height in Japan during the Muromachi period, other painting styles also flourished. Polychrome depictions of the patriarch reveal a consummate skill in execution. The eccentric visages of the disciples of the Buddha are found in a set attributed to the painter Ryōzen. They not only convey the persistent Zen fascination with spiritual force found in personality but also contain lush patterning and detail, as if a rugged eremitic type is slowly being enveloped in indigenous interests. These works, it should be said, also reflect dependence on Sung Chinese interpretations.



Arhat, hanging scroll attributed to Ryōzen, late 14th–early 15th century, Muromachi period. Ink and colour on silk. In the Freer Gallery of Art, Washington, D.C.

The Smithsonian Institution, Freer Gallery of Art, Washington, D.C.

The indigenous Yamato-e tradition also continued to develop during this period. The polished narrative painting forms found in the late Heian and Kamakura periods were still produced but were eclipsed by styles that conveyed energy at the expense of surface refinement. Their genesis paralleled the growth of narrative literature, which treated a growing number of legends and folktales. Also appearing with greater frequency was a narrative compositional technique that mixed word and image by juxtaposing text closely to the figure speaking the words, almost in cartoon style. Narrative painting

Screen painting in a rich polychromatic style persisted in parallel to the sparser, more obviously intellectual monochromes of the Zen tradition. The best of the Muromachi Yamato-e style screens show, in material and in sensibility, influences of metalworking, lacquering, and textile crafts. These works convey the reality of pragmatic creativity, which would come to full flower at the close of the 16th century.

The trends in Japanese calligraphy continued in essentially two major channels—the court-inspired, elegantly mannered script and the bold, ruggedly expressive forms of the Zen tradition.

Ceramics. The Muromachi-period taste in ceramics was, like painting, massively influenced by Chinese and Korean taste. Celadon ware was imported in large quantities. Known in Japan as Tenryūji ware, this light green monochrome ware was produced in many shapes as service ware and can be seen depicted in various narrative paintings of the period. It was imported as part of a large trading scheme managed by the Zen Tenryū Temple to support its works. Shogunal taste also favoured the sparse, darker ceramics from China, including *temmoku* ware, which revealed beautiful random effects in glaze colouring.

These comparatively austere Chinese ceramic types were gradually understood to have potential native equivalents in the ruggedly simple storage jars produced in Japanese kilns. Finely controlled glazes and enamel polychromy, which required the use of kaolin clay and controlled high firing, were still technically beyond Japanese capabilities; but the high regard in which the elegantly simple Chinese ware was held caused connoisseurs to elevate the status of once humble works and to commission Japanese interpretations of continental ware in Japanese kilns.

Lacquerware. Similarly, the appreciation of lacquerware was stimulated by the importation of fine Chinese works. The carved lacquer technique developed in Yüan China was emulated in a somewhat simpler manner in Japan. Lacquerware of a subdued red and black palette, said to have originated in the workshops of the Negoro Temple to the southeast of Ōsaka, was favoured in Buddhist establishments for its worn, unaffected look.

The tea ceremony. Perhaps the most calculatedly effective aesthetic development of the Muromachi period was the emergence of the cult of tea. The environment gradually required for tea gatherings grew into a kind of ritualized theatre in which objects removed from their original contexts were offered as worthy of consideration both in and of themselves and as metaphors for religious or philosophical perspectives. Zen monks imported tea plants from China, where the beverage was used for its medicinal qualities and as a stimulant in meditation. They also participated in a simple ceremony of consumption that included the use of certain prescribed utensils and implements.

From these fairly simple origins as a moment of respite and spiritual conviviality, the tea ceremony grew in complexity. Tea competitions (*tocha*) with the goal of discerning various blends began to be held in the Muromachi period and were espoused by Murata Shukō (*c.* 1422–1502), who was a disciple of the Zen master and abbot Ikkyū and is traditionally credited with founding the tea ceremony in Japan. An aesthetic adviser to the shogun Ashikaga Yoshimasa, Shukō prepared tea for his master at the latter's villa Ginkaku ("Silver Pavilion," now a temple) in a separate structure with a small tea room called Founder of the tea ceremony

the Dōjinsai. Shukō and those in his circle stressed the spiritual elements of the ceremony and encouraged the display of a piece of Zen calligraphy at the ceremony.

About this time the size of the tea ceremony room was standardized to four and a half tatami mats. Shelving, a recessed wall element or alcove (tokonoma), and other features provided places for displaying art appropriate to a season, mood, or other occasional intention. Implements such as tea cups, water jars, and kettles were carefully choreographed for the occasion.

The codification of the ceremony developed through the late Muromachi period and flowered in the succeeding Momoyama period. Similarly, the aesthetic intentions were more carefully articulated with time. In general, these goals included the cultivation of simplicity and the appreciation of rusticity. Within the careful ritual of tea preparation and sharing, the proper blend of object and participants was intended to heighten an awareness of transience and fragility. It trained the participant to be predisposed to learning from the simple and to seek new levels of meaning through the creative juxtaposition of objects, painting, and calligraphy. The practice of the tea ceremony had profound impact on the nature of fine art collecting by proposing new values for previously existing art and by encouraging the creation of works especially for use in the ceremony. In a time of radically shifting social alignments, it is noteworthy that the ambience of the tea ceremony thrived on suggested visual contrasts between the rustic and refined.

Architecture and garden design. The development of the tea ceremony encouraged architectural changes during the Muromachi period. The need for a small, discrete environment as a place of contemplation or connoisseurial consideration led to the evolution of both the tea room and a small study room, called *tsuke shoin,* containing a ledge used as a desk, shelves, and sliding shoji windows that opened onto an auspicious, usually man-made, view. The sprawling style of Heian-period construction, called *shinden-zukuri,* was modified to accommodate the reduced circumstances of the aesthete in the turbulent Muromachi period, and domestic architecture began to take on a more modest, carefully circumscribed, and mannered appearance.

The *shoin,* or study alcove

The consciousness of controlling an environment to produce effect was ever more evident and extended to the development of garden design. The various styles, whether dry or wet, presented a highly calculated series of meanderings and views. The prototypical aspiration of garden design was said to be an evocation of the environs of the Amida's Western Paradise.

Stately, symmetrical gardens, which reflected the ordered, aristocratic hierarchy and *shinden-zukuri* architectural style, are nowhere to be found in the Muromachi garden aesthetic. Retained from the tastes of previous periods was the penchant for blurring the line between created structure and nature; buildings were often constructed to be unpretentiously rustic, while gardens were meticulously designed to be viewed but not entered. Gardens were understood and meant to be read as a journey into a three-dimensional painting. The tea aesthetic was influential in their design. The careful reordering of nature in a "natural" way provided enlightened views for the careful observer. The hermitage and its natural surroundings became, in obviously mannered forms, an aesthetic touchstone for the times.

The Azuchi-Momoyama period. The brief span of time during which first Oda Nobunaga (1534–82) and then Toyotomi Hideyoshi (1536/37–98) began the process of unifying the warring provincial leaders under a central government is referred to as the Azuchi-Momoyama, or Momoyama, period.

The dating of the period is, like the name, somewhat relative. The initial date is often given as that of Nobunaga's entry into Kyōto in 1568 or as that of the expulsion of the last Ashikaga shogun, Yoshiaki, from Kyōto in 1573. The end of the period is sometimes dated to 1600, when Tokugawa Ieyasu's victory at Sekigahara established his hegemony; to 1603, when he became shogun; or to 1615, when he destroyed the Toyotomi family. It should be noted that the rigid application of an essentially political chronology to developments in the arts can be deceptive. Many important cultural figures were active not only during the Momoyama period but in the preceding Muromachi or succeeding Edo period as well. Similarly, artistic styles did not necessarily change with each change in political system.

Discrepancy between political and artistic periods

In any case, Nobunaga's rise is the referent event for the start of the period. He selected Azuchi, a town on the eastern shore of Lake Biwa, a few miles to the east of Kyōto, as the site of his new government. It was there that a purportedly magnificent castle (now known only through records) was constructed between 1576 and 1579 and destroyed shortly after Nobunaga's death. A product of military necessity as well as an extension of the bold and outsize personality of its resident, this innovative structure presented enormous decorative challenges and opportunities to Kanō Eitoku (1543–90), the premier painter of the period.

Nobunaga's successor, Toyotomi Hideyoshi, was, of the three hegemons of the period, perhaps the one most enthusiastically involved with the arts. He constructed several castles, including one at Momoyama, just to the south of Kyōto. The name Momoyama has since become associated, as has Azuchi, with the lavish and bold symbolizations of political power characteristic of the period.

The fact that the two castle sites lend their names to the era seems especially appropriate artistically because the castle was the single most important crucible for experimentation in the visual arts in the Azuchi-Momoyama period. The development of the castle also points up several salient features of the age: a display of massive power held by provincial warriors not previously noted for high cultural aspirations, growing confidence in national stability, and the conscription of artists to articulate the new mood.

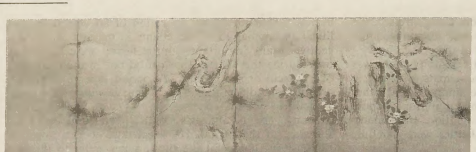
The development of the visual arts during this period was characterized by the vigorous patronage of two groups: the military leadership, who brought civil stability, and the merchant class, which formed the economic backbone of the revitalized urban centres. In addition, a much diminished aristocracy was still intent on retaining a hand in the arbitration of culture. Each group found not only genuine pleasure through their patronage of the arts but, in a time of major social realignments, legitimization and proclamation of their social status as well.

Architecture. The Japanese castle was a totally indigenous architectural form that developed in the 16th and early 17th centuries, a by-product of the hostile military conditions that existed in Japan from the time of the Ōnin War and in the following 100 years. Before that time military architecture had primarily consisted of small wooden fortresses; the earliest stone structure was probably constructed in the 1530s. Castle architecture experienced its florescence and most imaginative expression in the period from 1600 to 1615. In fact, most of the extant castles are products of that period. By 1615, however, each domain was allowed only one castle, and all other castles were ordered destroyed. Indeed, further castle building by the domain lords, or daimyo, was later banned. Continuing need for fortification would have implied either hostile intention or impending instability. Either suggestion was unacceptable to the Tokugawa rulers.

Castle layout

The general castle layout consisted of a donjon, or reinforced tower, called the *tenshu,* around which were arranged gardens, parks, and fortified buildings used for both official and private purposes. The whole was surrounded by deep moats and massive stone walls. Castle interiors presented a new dimension of decorative challenges. Large, generally dark spaces were subdivided by sliding panels (*fusuma*) and folding screens (*byōbu*). These two elements provided the format, depending on the wealth and predilection of the patron daimyo, for extensive painting programs. While architectural and religious iconographic needs of previous eras required paintings of considerable scale, the quantity, stylistic bravura, and thematic innovations of the Azuchi-Momoyama period are singular in the burst of confident national energy that they represent.

The *shoin* style noted first in the Muromachi period continued to be refined. A veranda linked the interior of most



"Pine and Camellias," one of a pair of six-panel screens attributed to Kaihō Yūshō, late 16th–early 17th century, Azuchi-Momoyama period. Ink, colour, and gold on paper. In the Cleveland Museum of Art. 157 × 358.4 cm.

© The Cleveland Museum of Art, John L. Severance Fund, 1987.41

structures with the carefully arranged, highly cultivated exterior gardens. An interior room with shelving and a tokonoma for the display of a hanging scroll of painting or calligraphy continued to be the primary showroom for fine arts, although *fusuma* and *byōbu* decorated with large-scale paintings could be found throughout the structure. The main room was often divided into two levels, the slightly raised one, which was backed by the tokonoma and fronted by decorative wood carving, being reserved for the highest-ranking person present. Unlike the *shinden* style, which used curtains and folding screens to partition small areas of a single large room, *shoin*-style structures were divided into several rooms by fixed walls and sliding doors. With variations in scale, this was also true for the architecture of religious establishments.

Painting. Painting was the visual art form that offered the most varied opportunities in the new age and, in fact, the most notable area of achievement. A breakdown of the comparatively rigid lines that had previously defined the various painting styles began in the Muromachi period and continued in the Momoyama. The Kanō school developed two distinctive styles: one featuring bright, opaque colours on gold or silver backgrounds, brilliantly amalgamating bright colour and bold brushwork, and the other a more freehanded, mannered, and bold interpretation of traditional ink monochrome themes. Other schools varied these two styles into distinctive lineage voices, but the Kanō group under Eitoku dominated the period through sheer talent and by amassing important commissions.

The Kanō school

At Eitoku's death several other figures who had worked either in secondary collaboration or in competition with the Kanō atelier emerged as strong individualist painters. Kaihō Yūshō (1533–1615) probably trained in the Kanō studio, but his independent style, most characteristically revealed in richly nuanced ink monochrome on gold or silver background, owed much to a careful study of Zen painting. Hasegawa Tōhaku (1539–1610) arrived in Kyōto from the Noto Peninsula region to the north on the Sea of Japan. His training was thoroughly eclectic, with experience in Buddhist polychrome themes, portraiture, and ink monochrome. Through the offices of the tea master Sen Rikyū, Tōhaku gained access to important collections of Chinese painting that had greatly influenced Muromachi aesthetics. His acknowledged masterworks are in both the full-blown but delicately nuanced polychrome style and the more subtle, contemplative ink monochrome format. The latter style is exemplified by the hauntingly depicted pine trees obscured by a mist that he painted on a pair of sixfold screens.

The subject matter favoured by the military patrons was bold and aggressive, as overtly suggested in paintings of birds of prey, lions, and tigers. Slightly more subtle but equally assertive renderings of majestic rocks or trees were also popular. Some Confucian themes, reflective of the ideology that would be favoured even more forcefully under Tokugawa rule, were beginning to appear. Yet another theme endorsed by rulers and townspeople was a style of genre painting that celebrated the new prosperity and stability, both urban and agrarian. Panoramic and carefully detailed screen paintings laid out the bustling life of Kyōto emerging from the destruction of civil-war life.

An aberrational but richly interesting thematic interlude involved the presence of Iberian merchants, diplomats, and missionaries. These Westerners were part of the vast exploration, trade, and colonization effort that reached South America, Africa, and South and Southeast Asia. From the time of the foreigners' first arrival in 1543 until their expulsion in the 1630s, there was a modest amount of cultural transmission. During this time the Japanese commissioned liturgical implements from the West and acquired some training in Western painting techniques. Perhaps most memorably, it became fashionable to depict Western themes and screen panoramas of the foreigners active in various Japanese settings—walking in the streets of Kyōto or arriving at ports in galleons. Unlike paintings with Japanese or Chinese themes, which are read from right to left, a telling curiosity of these screens is that they are read from left to right, suggesting by composition that the foreigners would depart. This exposure to the West seems to have had little long-term effect on the Japanese visual arts of the time.

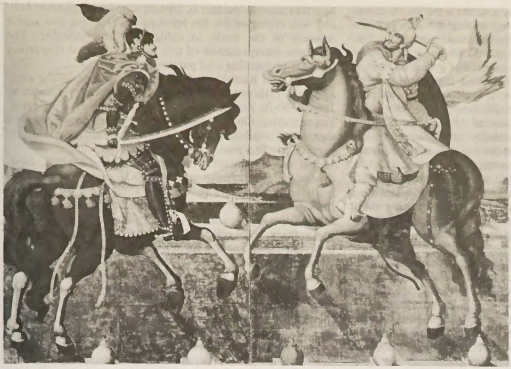
First introduction to Western culture

If the Kanō school and related interpreters advanced the themes and styles of the Muromachi period to accommodate the expansive sensibilities of the new ruling class and new social phenomena in general, yet another alignment of artistic talent offered a reexamination of the themes and expressive modes of the Heian court. The renaissance of courtly taste experimented with word and image, intermixing poetry, painting or design, lush decorative papers reminiscent of famous Heian secular and religious works, and countless narrative illustrations or allusive references to the *Tales of Ise* and to *The Tale of Genji.* It was during the late Momoyama and early Edo periods that a canonical body or stock of standardized referent classical illustrations began to coalesce.

The courtly themes were tackled by all schools but perhaps most effectively by the creative partnership of Hon'ami Kōetsu (1558–1637) and Tawaraya Sōtatsu (fl. 1600–30). Although, strictly speaking, they created most of their greatest works in the Edo period, Sōtatsu and Kōetsu developed their aesthetic sensibilities in Kyōto during the Momoyama period, and the inspiration for their later works can be found in the great creative freedom characteristic of that time.

Kōetsu and Sōtatsu

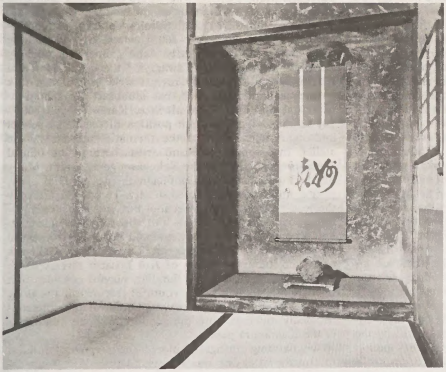
Kōetsu was raised in a family of sword experts, a discipline that required extensive knowledge of lacquer, metal, and leather. It implied an eye acutely attuned to delicate nuance in discerning the working of a blade. Kōetsu expanded his interests and training to include calligraphy and ceramics. He functioned as an impresario, bringing together talented craftsmen and artists to work on projects. None was more central to and intertwined with his reputation than Sōtatsu, a painter of fans. Both men, especially Kōetsu, had excellent connections with the aristocracy but came from artisan or merchant families. Working in collaboration, with Kōetsu acting as calligrapher, they created paintings and decorative backgrounds recalling the rich opaque texturing of an earlier illuminated sutra style.



Detail of "Xavier and the Western Princes on Horseback," two panels of a four-panel screen painting, Azuchi-Momoyama period. Colour and gold leaf on paper. In the Kōbe City Museum, Kōbe, Japan.
By courtesy of the Kobe City Museum

While both men, in other contexts, demonstrated mastery of the ink monochrome form, their works in polychromy featured a trait that would be characteristic of their followers throughout the Edo period: their images are formed through arrangements of colour patterns rather than being defined by ink outlines and embellished with colour. Ink was used more sparingly and allusively than, for example, by the Kanō painters. The effect was softening, textured, and suggestive of textile patterning. Sōtatsu's lush screen painting, said to describe the scene at Matsushima Bay on Japan's northeast Pacific coast, is a superb statement of elemental power couched in a decorative mode. Reference to the various planes of Chinese painting—near, middle, and far distance—were largely abandoned, as exposition of the surface of a material became the foremost concern.

Sōtatsu and Kōetsu worked in collaboration with the wealthy merchant Suminokura Soan (1571–1632), beginning in 1604, to produce images and calligraphy for a series of luxury-edition printed books featuring renderings of classical and nō drama texts. This collaboration marked the earliest and one of the most beautiful efforts at a wider dissemination of the Japanese classics to an increasingly literate audience. The energies and talents that these men and their followers infused into the Japanese visual arts were thoroughly unique. It may be suggested, however, that their initial training in art forms other than painting brought new pragmatism and perspective to the painting world.





The tea ceremony (cha-no-yu).
(Above) The Tai-an tea room in the Myōki-an, Kyōto, design attributed to Sen Rikyū, the great tea master of the Azuchi-Momoyama period. The floor is covered with tatami (straw mats), and in the rear is the tokonoma (alcove) where a wall hanging of calligraphy is displayed for the contemplation of the tea drinkers. (Right) Tea bowl called "Fujisan," gray raku ware by Kōetsu. In the Sakai Collection, Tokyo.
Photograph, (above) Sakamoto Photo Research Laboratory; by courtesy of (right) the International Society for Educational Information, Tokyo

Ceramics. The tea ceremony and the need for its attendant wares continued to develop during the Momoyama period. The ceremony itself enjoyed greater popularity, but the political instability of the late Muromachi and early Momoyama periods drove an important group of potters from Seto, near Nagoya, to the Mino region, somewhat northeast of their former site. It was in this area that many new and expanding commissions for tea ware were executed. Under the supervision of Mino kiln masters, subvarieties were produced, notably Shino ware, which used a rich feldspathic glaze whose random surface bursts and crackles appealed greatly to tea connoisseurs.

Raku ware

Works commissioned by the tea master Furuta Oribe (1544–1615) featured aberrant or irregular shapes, adding to the random effects of firing. In the Kyōto area raku ware was the characteristic type. This was typically a hand-shaped, low-fired, lead-glazed bowl form that had been immersed in cold water or straw immediately after being removed from the hot kiln in order to produce random, unique effects on the surface. In Kyushu, probably under the direction of Korean potters, a high-fired ceramic known as Karatsu ware was introduced in the early 1590s. The plain, unsophisticated shapes and designs of Karatsu ware made it especially popular for use in the tea ceremony. Tea wares featuring controlled peculiarities and manufactured rather than serendipitous defects were also introduced during the Momoyama period.

The Tokugawa, or Edo, period. At the death of the Momoyama leader Toyotomi Hideyoshi in 1598, his five-year-old son, Hideyori, inherited nominal rule, but true power was held by Hideyoshi's counselors, among whom Tokugawa Ieyasu (1543–1616) was the most prominent. Ieyasu assumed the title of shogun in 1603, and the de facto seat of government was moved from Kyōto to his headquarters in Edo (now Tokyo). Ieyasu completed his rise to power when he defeated the remaining Toyotomi forces in 1615. These events marked the beginning of more than 250 years of national unity, a period known as either Tokugawa, after the ruling clan, or Edo, after the new political centre.

The government system implemented by the Tokugawa rulers allowed for comparative discretionary rule within the several hundred domains, but the daimyo were required to pay periodic visits to Edo and to maintain a residence there in which family members or important colleagues remained, a gentle form of hostage holding and a major factor in the city's rapid growth.

In order to legitimize their rule and to maintain stability, the shoguns espoused a Neo-Confucian ideology that reinforced the social hierarchy placing warrior, peasant, artisan, and merchant in descending order. The early economy was based on agriculture, with rice as the measured unit of wealth. The warrior, the highest-ranking member of society, was salaried on rice and soon found his net worth fluctuating as wildly as the annual harvest yields. The merchant, on the other hand, who ranked lowest because he was understood to live off the labour of others, prospered in this time of peace and dramatic urban growth, a phenomenon that gave the lie to his theoretical value in the social order. Thus, from the inception of Tokugawa rule there was an implicit tension between the realities of a strong emerging urban culture, an inefficient agrarian economy, and a promulgated ideal of social order.

New cultural forms

The economic power of the merchant class and the expansion of the urban centres widened the audience for the arts from the traditional base of the nobility and the political elite. New cultural forms were generated, including the Kabuki theatre and the licensed brothel quarters. In a generally restrictive and controlled society, these entertainments served as a social safety valve offered by the shogunate to the merchant class (although participation in this world was egalitarian). Their popularity opened a whole new thematic source to the visual arts as the formats of woodblock print and painting were employed to depict the many facets of the pleasure quarters.

The shogunate's adaptation of Chinese concepts extended beyond Neo-Confucianism. China was again officially embraced as a source for models not only of good government but also of intellectual and aesthetic pursuits. The Chinese amateur scholar-painter (Chinese: *wen-jen,* Japanese: *bunjin*) was esteemed for his learning and culture and gentle mastery of the brush in calligraphy and painting. The Japanese interpretation of this model spawned important lineages of painting and patronage.

A final Zen Buddhist migration from China in the early and mid-17th century introduced the Ōbaku Zen sect to Japan. While not on the scale of Zen influence of previous centuries, Ōbaku monks provided the Japanese with a significant window on contemporaneous Chinese culture, particularly literature, calligraphy, and painting.

Most direct contact with foreigners was limited, however, especially after a policy of national seclusion was instituted in 1639. The Dutch trading post of Deshima in Nagasaki Harbour was Japan's primary window on the outside world, providing a steady stream of Western visual images, most often in print form and frequently once removed from Europe through a Chinese interpretation. Western themes, techniques, and certain optical technology suggested new ways of seeing to Japanese artists.

In the 19th century relations with the outside world ceased to be a controlled exercise in curiosity. Although Japan's limited natural resources offered no major temptation to colonizers, Western nations increased pressure on Japan to open its ports. The transition in sea travel from sail to steam put new demands on Western trading and naval fleets. Japan's strategic location, with its potential as a port for refueling and trade, was ever more evident. During the 1850s, treaties agreed to by a weakened shogunate raised the ire of many. In the south and west the domains of Chōshū and Satsuma, which held long-festering resentment of the Tokugawa reign, led rebellions during the 1860s. They overpowered the shogunal forces and "restored" the emperor in 1868, ending Tokugawa rule.

Nationalism

The feelings of nationalism that contributed to the imperial restoration had begun to develop in the 18th century, when a school of nativist ideology and learning arose. Partly in response to the shogunate's emulation of Chinese culture, this mode of thinking posited the uniqueness and inherent superiority of Japanese culture. In the visual arts this "national learning" (*kokugaku*) was expressed by an increase in an existing interest in courtly and classical themes.

Painting. The Kanō school of painters expanded and functioned as a kind of "official" Japanese painting academy. Many painters who would later begin their own stylistic lineages or function as independent and eclectic artists received their initial training in some Kanō atelier. Kanō Sanraku (1559–1635), whose bold patterning came closest among the early Kanō painters to touching the tastes stimulated by Tawaraya Sōtatsu and Hon'ami Kōetsu with their courtly revival style, provided a link to the generative energies that launched the school to its initial position of prominence. Kanō Tanyū (1602–74) solidified the dominant position of the Kanō school and significantly directed the thematic interests of the atelier. In a sense, the Kanō artists became the official visual propagandists of the Tokugawa government. Many of their works stressed Confucian themes of filial piety, justice, and correctly ordered society. Tanyū was not only the leading painter of the school but was also extremely influential as a connoisseur and theorist. Tanyū's notebooks containing his comments and sketches of observed paintings are a major historical source. His graceful ink and light colour rendering of Jizō Bosatsu reveals brush mastery and a thoroughly familiar, playful consideration of a Buddhist image. The youthful features of the deity are conveyed as at once fleshy and ethereal. The image is decidedly different from the gentle but stately renditions of the Kamakura period.

Two painting lineages explored the revival of interest in courtly taste: one was a consolidation of a group descending from Sōtatsu, and the other, the Tosa school, claimed descent from the imperial painting studios of the Heian times. The interpretations offered by the collaboration of Kōetsu and Sōtatsu in the late Momoyama period developed into a distinctive style called *rinpa,* an acronym linking the second syllable of the name of Ōgata Kōrin (1658–1716), the leading proponent of the style in

Rinpa



"The Bodhisattva Jizō Playing a Flute," hanging scroll by Kanō Tanyū, Edo period. Ink and light colour on paper. 99.9 × 38.9 cm. In the Mary and Jackson Burke Collection, New York City.
Mary and Jackson Burke Foundation, photograph by Carl Nardiello

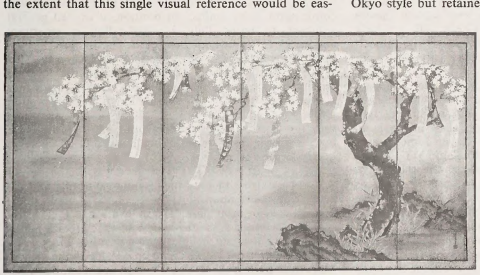
the Edo period, and *ha* (*pa*), meaning school or group. Sōtatsu himself was active into the 1640s, and his pupils carried on his distinctive rendering of patterned images of classical themes. Like Sōtatsu, Kōrin emerged from the Kyōto trades as the scion of a family of textile designers. His paintings are notable for an intensification of the flat design quality and abstract colour patterns explored by Sōtatsu and for a use of lavish materials. His homage to the Yatsu-hashi episode from the *Tales of Ise* is seen in a pair of screens featuring an iris marsh traversed by eight footbridges that is described in the story. Kōrin attempted this subject, with and without reference to the bridges, on several occasions and in other media, including lacquerwork. Classical literature had imbued popular culture to the extent that this single visual reference would be easily recognized by viewers of the period, permitting Kōrin to evoke a familiar mood or emotion without having to depict a specific plot incident. Other notable exponents of the *rinpa* style in the later years of the Edo period were Sakai Hōitsu (1761–1829) and Suzuki Kiitsu (1796–1858).

The Tosa school, a hereditary school of court painters, experienced a period of revival thanks to the exceptional talents and political acuity of Tosa Mitsuoki (1617–91). Mitsuoki's patronage connections to the imperial household, still residing in Kyōto, provided him with an appreciative aristocratic audience for his refined narrative evocations of Heian themes and styles. A pair of screens depicting spring-flowering cherry and autumn maple strike a melancholy chord. Attached to branches of the trees are decorated slips of paper bearing classical poems inscribed by the unseen participants in traditional court outings to celebrate the seasons. The allusion to past literary glory and to a poetry party recently dispersed suggests the mood of the court now resigned to ceremonial roles under the Tokugawa dictatorship. The Tosa atelier was active throughout the Edo period. An offshoot of the school, the Sumiyoshi painters Jokei (1599–1670) and his son Gukei (1631–1705), produced distinctive and sprightly renderings of classical subjects. In the first half of the 19th century, a group of painters, including Reizei Tamechika (1823–64), explored ancient painting sources and offered a revival of Yamato-e style.

In addition to the Kanō, *rinpa,* and Tosa styles of painting, which all originated in earlier periods, several new types of painting developed during the Edo period. These can be loosely classified into two categories: the individualist, or eccentric, style and the *bunjin-ga,* or literati painting. The individualist painters were influenced by nontraditional sources such as Western painting and scientific studies of nature, and they frequently employed unexpected themes or techniques to create unique works reflecting their often unconventional personalities.

Individualist painters

A lineage that formed under the genius of Maruyama Ōkyo (1733–95) might be summarily described as lyrical realism. Yet his penchant for nature studies, whether of flora and fauna or human anatomy, and his subtle incorporation of perspective and shading techniques learned from Western examples perhaps better qualify him to be noted as the first of the great eclectic painters. In addition to nurturing a talented group of students who continued his identifiable style into several succeeding generations, Ōkyo's studio also raised the incorrigible Nagasawa Rosetsu (1754–99), an individualist noted for instilling a haunting preternatural quality to his works, whether landscape, human, or animal studies. Yet another of Ōkyo's associates was Matsumura Goshun (1752–1811). Originally a follower of the literati painter and poet Yosa Buson (1716–83), Goshun, confounded by his master's death and other personal setbacks, joined with Ōkyo. Goshun's quick and witty brushwork adjusted to the softer, more polished Ōkyo style but retained an overall individuality. He and



"Flowering Cherry with Poem Slips," one of a pair of six-panel screens by Tosa Mitsuoki, c. 1675, Edo period. Ink, colour, gold leaf and powdered gold on silk. In the Art Institute of Chicago. 142.5 × 293.2 cm.
The Art Institute of Chicago, Kate S. Buckingham Collection, 1977.156, photograph by Robert Hashimoto © 1997, The Art Institute of Chicago, all rights reserved

his students are known as the Shijō school, for the street on which Goshun's studio was located, or, in recognition of Ōkyo's influence, as the Maruyama-Shijō school. Other notable individualists of the 18th century included Soga Shōhaku (1730–81), an essentially itinerant painter who was an eccentric interpreter of Chinese themes in figure and landscape conveyed in a frequently dark and foreboding mood. Itō Jakuchū (1716–1800), son of a prosperous Kyōto vegetable merchant, was an independent master of both ink and polychrome forms. His paintings in either mode often convey the rich, densely patterned texture of produce arrayed in a market.

Literati painting

The other new style of painting, *bunjin-ga,* is also called *nan-ga* ("southern painting") because it developed from the so-called Chinese Southern school of painting, which promoted the ideal of the learned scholar-gentleman who had no pecuniary or political interests and was unintimidated by the overly polished and spiritless examples of professional painting.

While the amateur ideal was pursued by many Japanese *bunjin,* the most remarkable of the ink monochrome or ink and light colour works were created by artists who, although generally attempting to conform to a *bunjin* lifestyle, were actually professionals in that they supported themselves by producing and selling their painting, poetry, and calligraphy. Especially notable artists from this tradition include the 18th-century masters Ike Taiga (1723–76) and Buson. Some of Taiga's most compelling works treat landscape themes and the melding of certain aspects of Western realism with the personal expressiveness characteristic of the Chinese *bunjin* ideal. Buson is remembered as both a distinguished poet and a painter. Frequently combining haiku and tersely brushed images, Buson offered the viewer jarring, highly allusive, and complementary readings of a complex emotional matrix. Uragami Gyokudō (1745–1820) achieved movements of near abstraction with shimmering, kinetic, personalized readings of nature. Tani Bunchō (1763–1840) produced paintings of great power in the Chinese mode but in a somewhat more polished and representational style. He was a marked individualist and served the shogun by applying his talents to topographical drawings used for national defense purposes. Bunchō's student Watanabe Kazan (1793–1841) was an official representing his daimyo in Edo. Through his interest in intellectual and artistic reform, he perhaps came the closest to exemplifying classic literati ideals. His accomplishments in portraiture are especially significant and reveal his keen study of Western techniques.

Ukiyo-e

Woodblock prints. A movement that paralleled and occasionally intersected with the aforementioned developments in painting was that of the production of ukiyo-e, or "pictures of the floating world," which depicted the buoyant, fleeting pleasures of the common people. This specialized area of visual representation was born in the late 16th and early 17th centuries as part of a widespread interest in representing aspects of burgeoning urban life. Depictions of the brothel quarters and Kabuki theatre dominated the subject matter of ukiyo-e until the early 19th century, when landscape and bird-and-flower subjects became popular. These subjects were represented in both painting and woodblock print form.

Woodblock printing had been a comparatively inexpensive method of reproducing image and text monopolized by the Buddhist establishment for purposes of proselytization since the 8th century. For more than 800 years no other single societal trend or movement had demonstrated a need for this relatively simple technology. Thus, in the first half of the 17th century, painters were the principal interpreters of the demimonde. The print format was used primarily for production of erotica and inexpensive illustrated novellas, reflecting the generally low regard in which print art was held. This perhaps resulted from the idea that the artist, when creating a painting, was essentially the producer and master of his own work. However, when engaged in woodblock print production, the artist was more accurately classified as the designer, who had been commissioned and was often directly supervised by the publisher, usually the impresario of a studio or other commercial enterprise.

Print production process

The simplest prints were made from ink monochrome drawings, on which the artist sometimes noted suggestions for colour. The design was transferred by a skilled carver to a cherry or boxwood block and carved in relief. A printer made impressions on paper from the inked block, and the individual prints could then be hand-coloured if desired. Printing in multiple colours required more blocks and a precise printing method so that registration would match exactly from block to block. Additional flourishes such as the use of mica, precious metals, and embossing further complicated the task. Thus, while the themes and images of the floating world varied little whether in painting or print, the production method for prints involved many more anonymous and critical talents than those of the artist-designer whose name was usually printed on the single sheet, and the mass-produced prints were considered relatively disposable despite the high level of artistry that was frequently achieved. Nevertheless, with the exponential increase in literacy in the early Edo period and with the vast new patronage for images of the floating world—a clientele and subject matter not previously serviced by any of the traditional ateliers—mass production was necessary, and new schools and new techniques responded to the market.

In the last quarter of the 17th century, bold ink monochrome prints with limited hand-colouring began to appear. "The Insistent Lover" by Sugimura Jihei (fl. *c.* 1681–1703) provides an excellent example of the lush and complex mood achievable with the medium. Within a seemingly uncomplicated composition Jihei represents a tipsy brothel guest lunging for a courtesan while an attendant averts her eyes. This scene, likely played out hundreds of times each evening in the urban licensed quarters, skillfully suggests the multileveled social games, including feigned shock and artful humouring of the insistent guest, that prevailed in the floating world. This print, too, with an almost naive representational quality, is an example of the generally straightforward, exuberant mood of the times in regard to the necessary indulgences.

Full-colour prints

From the late 17th until the mid-18th century, except for some stylistic changes and the addition of a few printed, rather than hand-applied, colours, print production remained basically unchanged. The technical capacity to produce full-colour, or polychrome, prints (*nishiki-e,* "brocade pictures") was known but so labour-intensive as to be uneconomical until the 1760s, when Suzuki Harunobu (1725?–70), whose patrons were within the shogun's circle, was commissioned to produce a so-called calendar print. Calendar manufacture was a government monopoly, but privately produced works were common. Seeking to avert any censorship, the private calendars were disguised within innocent-looking pictures. Harunobu's young woman rescuing a garment from the line as a shower bursts is an example of the technique. The ideograms for the year 1765 are part of the hanging kimono's pattern. More importantly, the work is a full-colour print. Even though it was commissioned for limited distribution, it excited general audiences to the possibilities of expanding the repertoire and appearance of woodblock prints. Harunobu's productions, through the end of the decade, elegantly suggested the new possibilities. His work so raised the level of consumer expectation that publishers began to enter full-colour production on the assumption that consumption levels would outweigh production costs. Not all prints were produced with the subtlety and care of Harunobu's, but the turn in taste toward full-colour prints, of whatever quality, was irreversible.

The last quarter of the 18th century was the heyday of the classic ukiyo-e themes: the fashionable beauty and the actor. Katsukawa Shunshō (1726–92) and his pupils dominated the actor print genre. His innovative images clearly portrayed actors not as interchangeable bodies with masks but as distinctive personalities whose postures and colourfully made-up faces were easily recognizable to the viewer. Masters at portraying feminine beauty included Torii Kiyonaga (1752–1815) and Kitagawa Utamaro (1753–1806). Both idealized the female form, observing it in virtually all its poses, casual and formal. Utamaro's bust portraits, while hardly meeting a Western definition of portraiture,

were remarkable in the emotional moods they conveyed. A mysterious artist active under the name of Tōshūsai Sharaku produced stunning actor images from 1794 to 1795, but little else is known of him.

At the close of the 18th century, a palpable tightening of government censorship control and perhaps a shift in public interest from the intense introspection provided by artists of the demimonde forced publishers to search for other subject matter. Landscape became a theme of increasing interest. In Edo the artist Katsushika Hokusai (1760–1849), who as a young man trained with Katsukawa Shunshō, broke with the atelier system and experimented successfully with new subjects and styles. In the 1820s and '30s, when he was already a man of some age, Hokusai created the hugely popular print series *Thirty-six Views of Mt. Fuji.* Andō Hiroshige (1797–1858) followed with another landscape-travelogue series, *Fifty-three Stations of the Tōkaidō,* which offered scenes of the towns and way stations on the central highway connecting Edo and Kyōto. Both these and other artists capitalized on public interest in scenes of distant places. These landscape prints in some way assuaged the restrictive travel codes enforced by the shogunate and allowed viewers imaginative journeys.

Hokusai and Hiroshige

Hokusai was also an important painter. His energetic rendering of the Thunder God is a fine example of the quirky and amusing quality of his figural painting. A characteristic swiftly modulating brush defines the figure, and light cast from an unseen source, perhaps lightning, allows for a play of light and shadow over the figure to model a sense of body volume.

Hiroshige painted as well, but his legacy is a vast number of prints celebrating scenes of a Japan soon to vanish. His "View from Komagata Temple near Azuma Bridge" is part of a series of 100 views of Edo. It demonstrates Hiroshige's finely honed abilities to effect atmosphere. The appearance of the cuckoo screeching in the sky alludes to classical poetry associated with late spring and early summer, as well as to unrequited love, while the tiny figures and the red flag of the cosmetics vendor suggest the transitory nature of life and beauty.

The depiction of famous views allowed for their idealization and also for important experiments with composition. Fragmentary foreground elements were used effectively to frame a distant view, a point of view adopted by some European painters after their study of 19th-century Japanese prints. Ironically, in their return to landscape and flora and fauna subjects, Japanese print arts revived the metaphoric vehicles of personal expression so familiar to the classic Japanese and Chinese painting traditions.

Although the time-tested themes of erotica, brothel, and theatre continued to be represented in 19th-century prints, an emerging taste for gothic and grotesque subjects found ample audiences as well. Historical themes were also popular, especially those that could be interpreted as critiques of contemporary politics. Ukiyo-e prints seemed to have been transformed from a celebration of pleasure to a means of widely distributing observations on social and political events. As the century closed, the print form, while active, was subsumed by the development of the newspaper illustration. This new form served many of the same purposes as prints and thus dramatically reduced the print audience, but it did not satisfy the needs of connoisseurs.

Ceramics. Ceramic and decorative arts flourished in the Edo period. While it was possible in almost all areas of the visual arts to accommodate the emergence of Japanese taste from the subdued or monochrome tastes of the Muromachi period to the burst of colour and pattern that was favoured in the Momoyama period, ceramics lagged behind. As the Edo period dawned, ceramic art also was able to participate in this development. Technological and supply limitations had previously hampered the ability of Japanese potters to produce a high-fired polychrome product. That problem was rectified in the early 17th century when the chambered climbing kiln was imported from Korea. This type of kiln was able to sustain controlled, high temperatures. Also, Korean potters working in western Japan discovered clay with a kaolin content high enough to allow vessels to be fired to the hard, fine surface classified as porcelain. In particular, white-glazed porcelain provided an excellent surface for the application of pigments to produce polychrome design patterns. Initial interest was in the imitation of Chinese blue-and-white ware, but the palette was quickly expanded.

Porcelain

The potter Sakaida Kakiemon (1596–1666), active in Arita in western Japan, was a pioneer in expanding the colour range and design patterns on the newly achieved creamy white surfaces. His works were especially admired in Europe. Also produced in the Arita region, by potters working for the Nabeshima clan, was a high-fired ware most frequently seen as footed shallow plates or dishes. The designs applied to the ware were typically bold and employed combinations of Yamato-e style painting and textile patterns. Kutani ware was produced in similar shapes, although denser designs and darker colours were used in the decoration. Kutani ware was primarily commissioned by the Maeda domain and, like Nabeshima ware, was not for public consumption or export.

Kyōto ceramics, already noted for the low-fired raku ware, responded to the fashion for porcelain with a break from the older traditions. Nonomura Ninsei (fl. 17th century) is the first identifiable Kyōto potter to use the high-fired, smooth-surfaced ware as a means to offer brilliantly coloured, painterly designs. Ninsei was far less interested than his predecessors in the inherent character of a vessel's randomly produced surface. His ceramic ware ranged beyond traditional vessels and included incense burners, candle holders, and other Buddhist liturgical implements. Kyōto artists who continued variations of the Ninsei legacy—referred to, after the place of production, as *kyōyaki*—included Ōgata Kōrin's brother Kenzan (1663–1743) and Aoki Mokubei (1767–1833). Kenzan's designs favoured uncomplicated and bold variations of *rinpa* painting style, while Mokubei's work reflected interest in Chinese sources.

Lacquerware. Throughout the Edo period, innovations in the production and design of lacquerware met the demands of a widening and appreciative audience. Extremely time-consuming to produce, the finished lacquer product is strong and resilient if properly handled. Ensembles for the study (such as writing tables and boxes to hold ink stones and brushes), furniture, and dining ware were among the most frequent uses. Paralleling in many ways the trends in ceramics, Edo-period lacquerware shifted from the sedate and simple styles of the Muromachi period as a taste for remarkably striking and complex objects developed. A delight in trompe-l'oeil was increasingly evident.

Lacquer was typically used for constructing *inrō,* the small, often multitiered and compartmentalized case that hung from a gentleman's kimono sash and held small belongings. It is perhaps in this format, especially from the late 18th century, that lacquer artists were most inspired by novelty and new fashions, such as the taste for verisimilitude. A lacquer *inrō* could be made to look as if it were made, for example, of aged and rotting wood or animal skin. Technically remarkable and frequently ingenious in construction and design, these and other objects were the aesthetic opposites of such early lacquer examples as Negoro ware. The later efforts seemed intent on disguising rather than revealing component materials.

Lacquer and *inrō*

Architecture and garden design. Architectural developments reflected the major tendencies found in other aspects of the visual arts. There were the quite differing perspectives provided by the aristocratic revival and the bombastic display favoured by the newly powerful. The mausoleum of Tokugawa Ieyasu, begun in 1636 and located in the mountainous area of Nikkō, north of Edo, features an abundance of polychrome decorative carving and exaggerated curving lines and is perhaps the quintessence of the floridly decorated, ostentatious form. But much residential architecture also began to feature elaborate decorative carvings on interior and exterior panels and joints.

The Katsura Imperial Villa, built between 1620 and 1624 on the southwestern edge of Kyōto, is the most outstanding example of a cohesive attempt to integrate a mannered interpretation of Heian styles with the architectural innovations spurred by the development of the tea ceremony. Carefully planned meandering paths lead to and from

the central structures through gardens dotted with small pavilion structures and tea huts offering orchestrated and allusive views. Perhaps a more moderate and quite beautiful example suggesting more subdued tastes within the shogunal and daimyo ranks is Kenroku Garden and its surrounding structures, located at Kanazawa, capital of the Maeda family domain on the Sea of Japan northeast of Kyōto. In general, the Edo garden, which underwent various refinements throughout the period, is bold and beautiful but more obviously crafted than the tea gardens of the Muromachi period. Nature's flaws have been disguised and the hand of the landscaper shows clearly.

Sculpture. Sculpture did not enjoy a great infusion of creativity during the Edo period. More obviously mannered and stylized interpretations of Buddhist deities and worthies were regularly produced. There were, of course, some sculptors of exceptional talent. Shōun Genkei (1648–1710) is renowned for his production (1688–95) of a set of 500 arhats (disciples of the Buddha) at Gohyaku Rakan Temple in Edo. His inspiration came from exposure to Chinese sculpture imported by Ōbaku Zen monks at Manpuku Temple to the south of Kyōto. Another expressive and thoroughly individualistic sculptor of the Edo period was the itinerant monk Enkū (1628?–95). He produced charming and rough-featured sculptures revealing bold chisel marks. His goals were to inspire faith and to proselytize. His works are totally without artifice, and the energy and power of his efforts are clearly conveyed.

Modern period. Japan's modern period is, for the purposes of this article, defined as beginning with the Meiji Restoration in 1868 and continuing through to the present. In the Japanese system of dating, this period encompasses the Meiji period (1868–1912), the Taishō period (1912–26), the Shōwa period (1926–89), and the Heisei period (1989–).

Modernity for Japan has been a process of seeking definition in its cultural and political relationships with other nations, both Asian and Western. Japan's official intentions toward the West during the Meiji period can be described as a calculated attempt to achieve Western industrial standards and to absorb Western culture at every possible level. Also during this time Japan was directly involved in two international conflicts: a war with China (1894–95) and a war with Russia (1904–05). Victorious in both these conflicts, Japan proved its ability to gear its newly established industrial base to the achievement of foreign expansionist goals. In 1910 Japan officially annexed Korea, a process it had begun in 1905 when it assumed a protectorate status over the peninsular nation. Japan's pretext was to establish a strong buffer zone against possible Western incursion, but Korea was essentially colonized as a source of labour and natural resources.

International relations

The Taishō period was characterized by a comparatively liberal mood both politically and in the arts. In retrospect it has been sometimes viewed as a romantic, euphoric period of cultural creativity following the more conservative Meiji era and preceding the militaristic mood of the 1930s. During this same period, as the Western powers with colonial and mercantile interests in Asia were forced to focus their attention on Europe during World War I (1914–18), Japan moved in to fill the vacuum, especially in China. The 1930s were characterized by a rise in militarism and further expansion on the Asian continent. This process culminated in World War II and in Japan's defeat by Western powers in 1945. The postwar period began with the Allied—almost exclusively American—occupation of Japan and was characterized by rebuilding, rapid growth and development, and increasing internationalism.

The development of the visual arts since 1868 has been considerably influenced by changing political climates and goals. Assuming an official posture of encyclopaedic investigation and selective assimilation of Western culture and technology in the late 19th century, the Japanese cultural mainstream was systematically infused with the classical forms of Western painting, sculpture, and architecture. In the first several decades of the Meiji period, there was an upheaval in patronage and in the status of the traditional Japanese artist. The Meiji government pursued a policy of officially separating whatever elements of Buddhism and Shintō had been joined over the centuries in Buddhism's attempt at a syncretic relationship with the indigenous religion. Buddhism was stripped of many tax privileges. There was, as well, a sometimes violent and destructive reaction to Buddhism owing to strong nativist sentiments. The cumulative effect of this official dismantling and unofficial persecution was to release a remarkable amount of Buddhist art onto the market. Temples were forced to divest in order to support themselves, and the patronage supplied to artists who created Buddhist iconography was seriously curtailed. Japanese artists also suffered because of the general trend to idealize all aspects of Western culture. Vast amounts of Japanese art, woodblock prints being the foremost example, went to Western collections.

This trend began to be reversed in the 1880s owing in part, ironically, to the efforts of Ernest Fenollosa (1853–1908), a recent Harvard graduate who arrived in Japan in 1878 as a philosophy instructor at the Tokyo Imperial University (now University of Tokyo). His avocational interest in Japanese and Chinese art became his passion. The Japanese reshuffling of cultural priorities placed him in a particularly advantageous position to acquire Japanese art—especially Buddhist art—of exceptional quality for Western collections. Working with a former student, Okakura Kakuzō (1863–1913; also known as Tenshin), Fenollosa also moved forcefully to influence the Japanese to reclaim their cultural heritage and to adapt creatively to changing tastes.

Influence of Ernest Fenollosa

The Japanese government sponsored the participation of Japanese artists and craftsmen in various late 19th- and early 20th-century international expositions held in Europe and America. These were of some help in advancing Western knowledge of Japanese culture. Collecting of the type endorsed by Fenollosa, as well as more popular collecting inspired by the flood of Japanese arts and crafts to Western markets, caused Westerners to take notice of a theretofore unknown visual arts tradition. This Western assessment of Japanese art was done in piecemeal fashion. Some obvious and immediate results included the influence of Japanese art on European and American painters and printmakers as well as the wider trend to Japanism. This initiation of communication between disparate visual worlds began the lengthy process of asserting the Japanese fine arts tradition within the context of world culture.

The early 20th century was not only a time of continued assimilation of Western art forms and philosophies but also a period in which traditional Japanese forms sought and achieved a new interpretive voice. With the rise of militarism, the visual arts were largely conscripted for straightforward propagandistic purposes or allowed only in thematically banal forms. Japan's defeat in World War II produced in many Japanese intellectuals and artists a distrust of the authority of the indigenous tradition, leading them to search for meaning in artistic movements and traditions abroad. Meanwhile, the postwar Allied occupation forces (1945–52) urged structural changes to ensure that Japan's cultural properties were properly honoured, protected, and made more widely available to general audiences. Censorship of contemporary materials, particularly for political content, continued to be imposed. In general, however, the occupation opened the way for international cross-cultural experimentation and the development of an "international style" that persists to the present.

Western-style painting. As early as 1855, preceding the Meiji Restoration, the Japanese established a bureau (later named Bansho Shirabesho, "Institute for the Study of Western Documents") to study Western painting as part of an effort to master Western technology. Technical drawing was emphasized in the curriculum. Takahashi Yuichi (1828–94), a graduate of that bureau, was the first Japanese artist of the period to express an artistic rather than strictly technical interest in oil painting. Through self-training and in consultation with the British illustrator Charles Wirgman (1835–91), then in Japan, his level of mastery increased. His "Still Life of Salmon" (1877), one of seven known attempts by Takahashi at the subject, elevates this ordinary subject to a splendid study of form and colour.

A school of fine arts was established in 1876, and a team

of Italian artists was hired to teach Western techniques. Most influential among them was Antonio Fontanesi (1818–81). Active as an instructor in Japan for only a year, Fontanesi, a painter of the Barbizon school, established an intensely loyal following among his Japanese students. His influence is seen in the works of Asai Chū (1856–1907), who later studied in Europe. Asai's contemporary Kuroda Seiki (1866–1924) studied in France under Raphael Collin and was among the most prominent exponents of a style that was strongly influenced by Impressionism in its informality and its use of lighter, brighter colours.

In general it can be observed that in the Meiji period there was an initial calculated strategy to study Western representational methods for the larger purpose of bringing Japan to a perceived level of modernity. However, a small but influential group of painters became involved in a cross-cultural exchange that could not be controlled by government planning. Oil paintings that vary in skill of execution from awkward to highly competent were produced during this time. Unlike the comparatively sympathetic modes of painting that Japan assimilated from China, Western painting posed conceptual as well as technical challenges. Not only did unfamiliar materials such as oil and canvas have to be mastered but also new theories of composition, shading, and perspective—and the underlying Western philosophy of nature and its representation that had led to their development over the centuries—had to be absorbed.

The close of the Meiji period saw greater rigidification of painting schools, affiliations, and systems of official recognition through annual exhibitions. There were government-sponsored exhibitions and associations as well as protest salon or secessionist groups indicating a lively spirit of resistance to official control.

An increasingly sophisticated understanding of European art and cultural trends marked the Taishō period. The humanistic literary journal *Shirakaba* ("White Birch") was devoted to and highly influential on these subjects, and it was instrumental in introducing Japanese artists to European Impressionism and Postimpressionism. Its publication period (1910–23) essentially spanned the Taishō era. The paintings of Kishida Ryūsei (1891–1929) exemplify the extensive assimilation of sympathetic European moods into a Japanese mode. Kishida was a devoted follower of the Dutch painter Vincent van Gogh and later of artists of the Northern Renaissance such as Albrecht Dürer and Jan van Eyck. "Reiko with a Woolen Shawl" (1921), Kishida's portrait of his six-year-old daughter, attributes a knowing maturity to the sitter, an effect achieved, in part, through a slight distortion of features to produce a gnomelike adult visage.

The conscription of efforts during the war years enforced on the visual arts choices of severe puritanism, blithe optimism, or heroism. The work of Umehara Ryūzaburō (1888–1986) is a case in point. In the early 20th century he studied with Asai Chū and in France with Pierre-Auguste Renoir. His ebullient palette and love of patterning, as seen in his famous "Tzu-chin-ch'eng Palace" (1940), convey a cheerful mood. Not revealed in the painting is its occasion: the artist is present in Peking (Beijing) as part of an occupying force in the midst of war.

In the postwar period Japanese artists enjoyed widely increased access to Western sources and the ability to travel more freely in the West. As a result, practitioners of virtually all modern artistic movements and styles—including abstract expressionism, minimal and kinetic art, op and pop art—can be found in Japan.

Japanese-style painting. Paralleling the intensive and systematic study of Western painting methods was a steady process of renewal occurring in the field of traditional painting. Fenollosa was particularly instrumental in redirecting and salvaging the careers of two important late 19th-century painters, Kanō Hōgai (1828–88) and Hashimoto Gahō (1835–1908). Fenollosa had particular notions about the ways these traditional Kanō-school painters could adapt their techniques in order to create a more exciting and, perhaps to Western eyes, a more marketable product. He encouraged the use of chiaroscuro, brilliant palettes, Western spatial perspective, and dramatic atmospherics, and these techniques were indeed effective in creating new interest in the previously moribund forms of traditional Kanō painting.

A generation of painters inspired by the success of Hōgai and Gahō sought to expand the technical adaptations of these masters. Shimomura Kanzan (1873–1930), Yokoyama Taikan (1868–1958), and Hishida Shunsō (1874–1911) stand at the beginning of the *nihonga* ("Japanese painting") movement, in which traditional Japanese pigments were used but with a thematic repertoire much expanded. Format was no longer limited to scroll or screen and included occasional Western framed paintings. Shimomura's portrait of Okakura Kakuzō pays homage to Okakura's role as a mentor to the movement. This is a preparatory sketch for a completed portrait unfortunately destroyed in the Great Kantō Earthquake of 1923. Yokoyama and Hishida sought out more international, often Asian, themes. Their *nihonga* used the materials of traditional Yamato-e painting but, like the later Kanō paintings, incorporated heightened dramatic and atmospheric effects.

The *nihonga* movement

Maeda Seison (1885–1977), prominent in the next generation of *nihonga* artists, which also included Imamura Shikō (1880–1916), Yasuda Yukihiko (1884–1978), Kobayashi Kokei (1883–1957), and Hayami Gyoshū (1894–1935), employed an eclectic assortment of earlier Japanese painting techniques. At Okakura's suggestion he studied *rinpa.* His use of *tarashikomi,* a classic *rinpa* technique that achieves shading through pooling successive layers of partially dried pigment, clearly points out his wide-ranging adaptation of traditional techniques. Seison and others of his period were especially fond of historical subjects.

A somewhat distinct tradition of *nihonga* developed in Kyōto, finding natural precedents in the lyrical realism of the Maruyama-Shijō school of painters. Takeuchi Seihō (1864–1942) was the most successful proponent of this lineage. Interestingly, his most distinguished student was Uemura Shōen (1875–1949), a woman who revived a style reminiscent of ukiyo-e beauty portraits but instead idealized women in domestic settings.

Nihonga continued to flourish after World War II. This essentially traditional style was energized, like other Japanese art forms, by the openness of the postwar years. Traditional themes of flora, fauna, and landscape were joined by abstractions and by modern urban and industrial scenes. The resulting works, which use traditional pigments and brushes, provide a curious Japanese version of modernism.

The literati movement seemed to proceed with the least disruption of any of the traditional lineages. Tomioka Tessai (1837–1924) stands out as perhaps the latest of the literati masters. Noted for dense, rough brushwork and occasionally jarring choices in bright pigments, his creations were animated, cheerful evocations of Sung dynasty poetry.

Woodblock prints. The world of woodblock prints was profoundly affected by the changes ushered in during the Meiji period. The print medium had long served both connoisseur and general audience. With the advent of mass-circulation newspapers, however, the latter group was coopted. Illustrators and designers produced reportorial images and cartoons for newspapers, satisfying the public demand for illustration but removing a large block of economic support from the traditional print publishers. Print artists nevertheless continued to document the remarkably varied moods of the period. For example, a type of print known as Yokohama-e, named after the Japanese port city with a large resident foreign population, offered glimpses of the customs and appearances of the recently arrived visitors. Brutal, grotesque, and dark-humoured visions by such artists as Kawanabe Gyōsai (1831–89) and Tsukioka Yoshitoshi (1839–92) suggested that assimilation with the West was a socially and psychically traumatic process. Kobayashi Kiyochika (1847–1915), a student of Charles Wirgman as well as of Gyōsai, is best known for his prints illustrating the Sino-Japanese War and for his highly successful visions of contemporary Tokyo.

In the early 20th century two general currents moved the print world. The *shin hanga* ("new print") movement

The *shin hanga* movement

sought to revive the classic ukiyo-e prints in a contemporary and highly romanticized mode. Landscapes and women were the primary subjects. Watanabe Shōsaburō (1885–1962) was the publisher most active in this movement. His contributing artists included Kawase Hasui (1883–1957), Hashiguchi Goyō (1880–1921), Yoshida Hiroshi (1876–1950), and Itō Shinsui (1898–1972). Hashiguchi was determined to have complete control over his artistic output, and his tenure as a Watanabe artist was brief. His prints numbered only 16 and were mostly studies of Taishō women in a fashion thoroughly reminiscent, in technique and in composition, of Utamaro.

The other woodblock print trend was *sōsaku hanga,* or "creative print," a movement modeled on European approaches to print production. The artist, instead of consigning his designs to the carvers and printers employed by the publisher, performed all aspects of production. This was a philosophy of total engagement with the work. The leader of this movement was Onchi Kōshirō (1891–1955). Also prominent was Yamamoto Kanae (1882–1946). A notable feature of *sōsaku hanga* works was a movement toward defining shapes using colour rather than outlines, as in traditional woodblock prints.



"On Deck," woodblock print by Yamamoto Kanae, *c.* 1917. In the Art Institute of Chicago. 15.5 × 18.5 cm.
The Art Institute of Chicago, Clarence Buckingham Collection, 1979.633, photograph © 1997, The Art Institute of Chicago, all rights reserved

The print medium continues to be a particularly fertile arena of development in the Japanese visual arts. The use of the woodblock print has largely been usurped by lithography and other techniques, although there are periodic resurgences of interest in woodblock. Themes vary widely from traditional representational to abstract. The relatively inexpensive and easily portable format has made the modern Japanese print, and thus Japanese visions of modernity, widely available to international collectors.

Sculpture. Sculpture in the modern period was most productive in the bronze medium. The Italian Vincenzo Ragusa, along with other foreign technical experts recruited in the late 1870s, was a major influence in instructing young Japanese artists in bronze casting, although he privately despaired of their abilities at three-dimensional conceptualization. Japanese sculptors applied the new format to nonreligious subjects, including portraits and studies of anonymous subjects in a celebration of Japanese physical types. Takamura Kōtarō (1883–1956) was particularly influenced by Auguste Rodin, as was Ogiwara Morie (1879–1910), who produced notably fine heroic figures.

In the postwar period, Japanese sculptors and their works became more visible at international art fairs and competitions. As in other media, traditional formats fell from favour. Abstract forms have dominated the contemporary sculptural field, which has also been marked by experimentation with diverse materials. Installation art has joined the larger sculptural repertoire, and outdoor sitings—both in open natural spaces and in urban environments—have attracted much interest. Massive creations in bamboo and other works that interact with the environment are especially popular.

Ceramics. In addition to the continuation of various traditional lineages, the most significant development in ceramics of the modern period was the return to folkcraft tastes. Yanagi Sōetsu (1889–1961) espoused anonymity, functionality, and simplicity as a corrective to the industrialism and self-aggrandizement characteristic of the age. In league with potters such as the British artist Bernard Leach (1887–1979), Hamada Shōji (1894–1978), and Kawai Kanjirō (1890–1966), Yanagi engendered a robust, charming type of ceramic which recalled the wares that appealed to tea masters of the Muromachi and Momoyama periods. Kitaōji Rosanjin (1883–1959) was the major exponent of highly decorated work in the Kutani and later *kyōyaki* traditions. His role was largely as designer and production manager. Long associated with a famous restaurant, he was most conscious of the choreography of a total sensory experience in which his wares were an essential element.

Contemporary Japanese ceramics follows both traditional and abstract lines. Developments have been marked by wide experimentation in form and a general movement from traditional, functional pieces to "art" or sculptural works. The line between sculptor and ceramicist has become increasingly blurred.

Architecture. Japanese architecture created from the last quarter of the 19th century is remarkable in its rapid assimilation of Western architectural forms and the structural technology necessary to achieve results quite foreign to traditional Japanese sensibilities. Large-scale official and public buildings were no longer constructed of wood but of reinforced brick, sometimes faced with stone, in European styles. Steel-reinforced concrete was introduced in the Taishō period, allowing for larger interior spaces.

English and German architectural influences

The English architect and designer Josiah Conder (1852–1920) arrived in Japan in 1877. His eclectic tastes included adaptations of a number of European styles, and the work of his Japanese students was significant through the second decade of the 20th century. The Bank of Japan (1890–96) and Tokyo Station (1914), designed by Tatsuno Kingo (1854–1919), and the Hyōkeikan (1901–09), now an archaeological museum within the complex of buildings at the Tokyo National Museum, and the Akasaka Detached Palace (1909), both by Katayama Tōkuma (1853–1917), are but a few of the best-known examples of Japanese attempts at stately monumentality in a Western mode.

The German architects Hermann Ende and Wilhelm Böckmann were active in Japan from the late 1880s. Their expertise in the construction of government ministry buildings was applied to the growing complex of such structures in the Kasumigaseki area of Tokyo. The now much-altered Ministry of Justice building (1895) is a major monument to their work. The Germans also trained a group of protégés, including Tsumaki Yorinaka (1859–1916). His design of the Nippon Kangyō Bank (1899; no longer extant) and Okada Shinichirō's (1883–1932) Kabuki Theatre (1924) in Tokyo are representative of attempts to combine the grand scale of Western buildings with such traditional elements of Japanese architecture as tiled hip-gabled roofs, curved Chinese gables, and curved, overhanging eaves.

The striving for monumentality reached its most awkward form in the highly nationalistic period of the 1930s. The Tokyo National Museum (1937) by Watanabe Hitoshi and the Diet Building (1936), Tokyo, designed by Watanabe Fukuzo are examples of massive, blocky scale without grandeur.

Frank Lloyd Wright's Imperial Hotel in Tokyo (1915–22; dismantled in 1967) seemed to have little lasting influence, although Wright's creations in the West revealed his indebtedness to his perceptions of the Japanese aesthetic. Similarly, the Bauhaus movement stirred interest in Japan, but Walter Gropius was even more thoroughly impressed and influenced by such Japanese classics as the Katsura Imperial Villa in Kyōto.

Postwar architecture, while widely eclectic and international in scope, has seen its most dramatic achievements in contemporary interpretations of traditional forms. The structures created for the 1964 Tokyo Olympics by Tange Kenzō (b. 1913) evoke early agricultural and Shintō architectural forms while retaining refreshing abstraction. The residential and institutional projects of Andō Tadao (b. 1941) are marked by stark, natural materials and a careful integration of building with nature. In general, Japanese architects of the 20th century have been fully conversant in Western styles and active in developing a meaningful modern style appropriate to Japanese sites. (J.T.U.)

MUSIC

A study of the music of East Asia covers historical periods of changing styles from at least 2,000 BC to the present time. After a general introduction, the development of the musical systems of China, Korea, and Japan will be reviewed separately in chronological order.

The nature of East Asian music

EAST ASIAN MUSIC VIS-À-VIS THAT OF OTHER MAJOR CULTURES

East Asia can be viewed as one of the big four among the generally urban, literate cultural areas of the world. The other three are South Asia, the Middle East, and Europe. Around each of these major regional cultures one can find many satellite musical systems known as national forms. In most cases, the fundamental musical concepts of such national forms reflect the basic ideals of the cultural core.

Intercultural influences

Using instrument type alone as a measure, it is sometimes possible to note cultural influences and mixtures of the major traditions in smaller units. For example, the physical structure and playing positions of various bowed instruments in mainland Southeast Asia can often mark clearly Chinese influence, as in Vietnam, or Muslim and Chinese forms in confluence, as in the various bowed lutes of courtly ensembles in Cambodia and Thailand.

Concepts of music. If one turns to distinctions in musical style, one of the first questions to arise is "What is music?" Two basic definitions will suffice for the present discussion. The first definition is cultural: a sonic event can be called music if the people who use it call it music, regardless of one's own reaction to it. Similarly, certain events that sound musical to foreign ears are not music culturally if they are not accepted as such by native culture carriers. A good example of such a situation is found in the Middle East, where singing is never allowed in the mosque, though one may hear performances and even buy records of "readings" from the Qur'ān. Such cultural and functional problems of definition seldom arise in East Asian music, and a more neutral definition is appropriate. A sound event may be considered and studied as music if it combines the elements of pitch, rhythm, and loudness in such a way that they communicate emotionally, aesthetically, or functionally on the levels that either transcend or are unrelated to speech communication. Those who have been moved by a love song or a lament can well appreciate some of the implications of such a view of music. When listening to "exotic" music—*i.e.,* that of a tradition outside one's own background—it is important to remember that such transcendental values are at work for the alien listener as well as for listeners familiar with the particular musical language in use.

There are many kinds of music in the world, the three most common terms being folk, popular, and art music. Folk and popular music have their special indigenous and mixed forms in Asia (as in all the world today), but it is in the literate art traditions of Asia that historical and musical distinctions can be made most clearly. In the context of this discussion, art music is defined as a tradition having, to some degree, a conscious theoretical basis and a sense of repertoire that is played against the highest standards held by informed native listeners. The performer is often a professional, and there may be a known historical depth to the traditions. Thus, there may be art music in many nonliterate cultures such as that of the Australian Aborigines and that of the tribal courts of Africa. Here, however, the major concern is with one of the large urban, literate cultures and its three national variants. Before looking at these musical systems in detail, it is useful to compare the entire culture with those of the other major "big" three, South Asia, the Middle East, and Europe.

Theoretical systems. All four major literate cultures, in their ancient forms, laid a strong emphasis on the extramusical qualities of music. For example, the study of such concepts as the power of vibrations (in ancient Indian music theory) and the relationships between music and other elements in the universe (in Assyrian records as well as in the writings of medieval European scholars) can be matched in East Asia by the joint efforts of Chinese musicologists and astrologers to bring the music of the empire in tune with the universe.

In addition, all four cultures developed mathematically and acoustically based music theories. The pitches produced by dividing the length of a string were the basis of the three non-East Asian music theories. String acoustics were known in China as well, but, as described below, East Asian writings use the overtones of end-blown bamboo tubes to illustrate their systems. It is fascinating that, whatever their origin, the Middle Eastern and South Asian theories produced highly variable tone systems while the two ends of the old continent (*i.e.,* the West and China) generated 12 tones based on a cycle of pitches 5 tones apart (such as C to G to D in the West). This cycle of fifths produced 12 pitches that were mathematically correct, but the 13th pitch did not match the 1st pitch. In the West this so-called "Pythagorean comma" became bothersome as Western music oriented toward vertical sounds called harmony in which the distance between pitches in chords needed to be the same in every key. In the 17th century Western acousticians developed a formula that allowed them to bypass the "natural" tone system by making all pitches equidistant. The same formula was discovered by a Chinese mathematician and musicologist, Chu Tsai-yü, in the late 16th century; but such "well-tempered" tuning was not accepted in Chinese music practice until very recently, when Western music styles combined with indigenous traditions. This is one reason why Asian music sometimes sounds "out of tune" to Western ears.

The scientific base of music is reshaped by each culture into a system that meets its needs and tastes. There are differences in the sound, the instrumentation, and the forms of Western and Eastern music. Many will be noted in the subsequent chronological study of East Asian music. However, if the wonder of such variants is to be fully appreciated, it must be understood that music is not in fact an international language. It consists of a whole series of equally logical but sometimes very different closed systems. The word closed is used to mean that the musical facets mesh perfectly within a given system, but they often may prove difficult or impossible to transfer to another system. In this light, a given passage of Chinese music when analyzed or judged with the logic of Beethoven is chaos, but Beethoven seems equally illogical when viewed in the context of Chinese, or for that matter Indian, music theory. Such intercultural clashes can be constructed between almost any of the larger systems. In this context, one can see that Chinese music is tonally more foreign to Middle Eastern or Indian music than to Western, though historically it had closer relations with the other two. There are, of course, many other musical concepts and styles that traveled over the Silk Road between China and other parts of Asia; but these must be held aside until the discussion of the history of Chinese music.

Internally logical, closed systems

MUSICAL TRAITS COMMON TO EAST ASIAN CULTURES

In these primary considerations a view of some general aesthetic traditions common to much of East Asian music

is also requisite. The tonal vocabulary of 12 tones generated in a cycle of fifths is the first common factor. From this tonal vocabulary various scales of five to seven notes are chosen. Specific examples are found below in each area study. As in the West, the total number of notes in an East Asian scale is often seven, but each scale tends to have what could be called a five-tone (pentatonic) core (see notations III, VIII, and IX below). The one scale in which no half steps appear (the so-called anhemitonic pentatonic) is common all over the world, although casual listeners often mistake it as being characteristically "Oriental." A study of the scales found in this article or a few moments spent listening to authentic East Asian recordings will reveal clearly that the five black notes on the piano do not represent all the sonic resources of Asian music. Indeed, there are a great variety of East Asian musics. Their three most common characteristics are linearity, transparency, and word orientation.

Linearity

Linearity means an emphasis on melodic tension and release supported by or held in further tension by rhythmic devices. This line-and-rhythm orientation and lack of interest in Western-style harmony are, in fact, major distinctions between most of the world's music and that of the West. In traditional East Asian music, as well as in most other non-Western traditions, all melodic instruments play the same basic melody. No one fills in the texture with chords. If harmonic texture is used, its function is to provide colour rather than to generate tension or release by chord progression. Heterophony (more than one version of the melody being heard at the same time) may occur to enrich the line. The sense of moving through a time continuum toward an ending, however, is basically developed through the tension produced during the wait for a pitch to resolve to a pitch of rest just one tone above or below.

Transparency refers to the preference in East Asian music for chamber-music sound ideals; no matter how large an ensemble may be, the individual instruments are meant to be heard. This differs from the orchestral sound ideal, popular in 19th-century Western music, in which the intention is to merge the sounds of the individual instruments into one musical colour. A transparent texture is a logical choice for a tradition that wishes to emphasize lines; orchestral colour helps to merge various lines into single vertical sonic events called harmony.

Word orientation refers to the fact that until the 20th century there was little abstract instrumental music, such as a sonata or a concerto, in East Asia. A piece had either a text or a title that evoked an image, such as "Moon over the River" or "Spring Sea." Perhaps this relates to a general sensitivity to nature in East Asian culture as a whole. Whatever its source, it has produced many sonorous and pleasing results.

The music of China

One always approaches any survey of Chinese music history with a certain sense of awe—for what can one say about the music of a varied, still active civilization whose archaeological resources go back to 3000 BC and whose own extensive written documents refer to endless different forms of music in connection with folk festivals and religious events as well as in the courts of hundreds of emperors and princes in dozens of different provinces, dynasties, and periods? If a survey is carried forward from 3000 BC, it becomes clear that the last little segment of material, from the Sung dynasty (AD 960–1279) to today, is equivalent to the entire major history of European music. For all the richness of detail in Chinese sources, it is only for this last segment that there is information about the actual music itself. Yet the historical, cultural, instrumental, and theoretical materials of earlier times are equally informative and fascinating. This mass of information will be organized into four large chronological units: (1) the formative period, from 3000 BC through the 4th century AD, (2) the international period, from the 4th through the 9th century, (3) the national period, from the 9th through the 19th century, and (4) the world music period of the 20th century.

FORMATIVE PERIOD

Ancient artifacts and writings. Chinese writings claim that in 2697 BC the emperor Huang-ti sent a scholar, Ling Lun, to the western mountain area to cut bamboo pipes that could emit sounds matching the call of the phoenix bird, making possible the creation of music properly pitched for harmony between his reign and the universe. Even this charming symbolic birth of music dates far too late to aid in discovering the melodies and instrumental sounds accompanying the rituals and burials that occurred before the first historically verified dynasty, the Shang (mid-16th to mid-11th century BC). The beautiful sounds of music are evanescent, and before the invention of recordings they disappeared at the end of a performance. The remains of China's most ancient music are found only in those few instruments made of sturdy material. Archaeological digs have uncovered globular clay ocarinas (*hsün*), tuned stone chimes (*ch'ing*), and bronze bells (*chung*); and the word *ku,* for drum, is found incised on Shang oracle bones.

Chou dynasty musical documents

The earliest surviving written records are from the next dynasty, the Chou (1111–255 BC). Within the famous Five Classics of that period, it is in the *Li chi* ("Record of Rites") of the 2nd century BC that one finds an extensive discussion of music. The *I Ching* ("Classic of Changes") is a diviner's handbook built around geometric patterns, cosmology, and magic numbers that indirectly may relate to music. The *Ch'un-ch'iu* ("Spring and Autumn") annals, with its records of major events, and the *Shu Ching* ("Classic of History"), with its mixture of documents and forgeries, contain many references to the use of music, particularly at court activities. There are occasional comments about the singing of peasant groups, which is an item that is rare even in the early historical materials of Europe. The *Shih Ching* ("Classic of Poetry") is of equal interest, for it consists of the texts of 305 songs that are dated from the 10th to the 7th centuries BC. Their great variety of topics (love, ritual, political satire, etc.) reflect a viable vocal musical tradition quite understandable to modern radio or record listeners. The songs also include references to less durable musical relics such as the flutes, mouth organ (*sheng*), and, apparently, two forms of the zither (the *ch'in* and the *se*).

Aesthetic principles and extramusical associations. Despite the controversial authenticity and dates of ancient Chinese written sources, a combined study of them produces tantalizing images of courtly parties, military parades, and folk festivals; but it does not provide a single note of music. Nevertheless, in keeping with the prehistoric traditions of China, the philosophies of sages, such as Confucius (K'ung-fu-tzu, 551–479 BC) and Mencius (Meng-tzu, *c.* 371–*c.* 289 BC), and the endless scientific curiosity of Chinese acousticians furnish a great deal of rather specific music theory as well as varied aesthetic principles. The straightest path to this material is found on the legendary journey, mentioned earlier, of Ling Lun in search of bamboo pipes. The charm of such a tale tends to cloud several interesting facts it contains. First, it is noteworthy that the goal of the search was to put music in tune with the universe. This extramusical need was noted earlier in the general discussion of world music history. It is upheld in theory in the "Annotations on Music" ("Yüeh-chi") section of the *Li chi* with such comments as "Music is the harmony of heaven and earth while rites are the measurement of heaven and earth. Through harmony all things are made known, through measure all things are properly classified. Music comes from heaven, rites are shaped by earthly designs." Referring back to the previously given general definition of music, it can be seen that such cosmological ideals may be not merely ancient superstitions but actually cogent insights into the cultural function of music in human societies. Confucius, as pictured in *The Analects* written long after his death, had a similar view of music, including a concern for the choice of music and modes proper for the moral well-being of a gentleman. It is an open question as to how much performance practice followed the admonitions and theories of the scholars; but centuries later one finds numerous pictures of the wise man standing before some

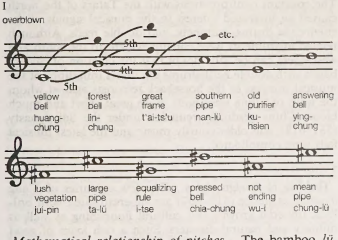
natural beauties while his servant follows closely behind him carrying his seven-stringed zither (*ch'in*) for proper use in such a proper setting.

Another point to be noted in the legend of the origin of music is the fact that Ling Lun went to the western border area of China to find the correct bamboo. It shall be noted as this article progresses how often cultures from Central and West Asia or tribal China influenced the growth and change of music in Imperial China. Finally, it is significant that, although the emperor in the myth was primarily concerned with locating pipes that would bring his reign into harmony with the universe, the goal was also the creation of precise, standard pitches.

Tonal system and its theoretical rationalization. As noted earlier, harmonic pitches produced by the division of strings were known in China. They may have been used to tune sets of bells or stone chimes, but the classical writings on music discuss a 12-tone system in relation to the blowing of bamboo pipes (*lü*). The first pipe produces a basic pitch called yellow bell (*huang-chung*). This concept is of special interest because it is the world's oldest information on a tone system concerned with very specific pitches as well as the intervals between them. The precise number of vibrations per second that created the yellow bell pitch is open to controversy (between middle C-sharp (C♯) and the F above) because the location of this pitch could be changed by the work of new astrologers and acousticians on behalf of a new emperor, in order that his kingdom might stay in tune with the universe. (The note C is used in notation I in deference to Western readers; it should not be assumed that a pitch identical to C♯ was necessarily central to ancient Chinese music). The choice of the primary pitch in China had extramusical as well as practical applications, for the length of the yellow bell pipe became the standard measure (like a metre); and the number of grains of rice that would fill it were used for a weight measure. Thus, the pipe itself was often the property not of the Imperial music department but of the office of weights and measurements.

Practical applications of Chinese music

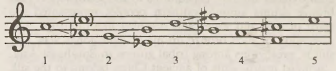
I



Mathematical relationship of pitches. The bamboo *lü* pipe is closed at the bottom by a node in the bamboo, with the result that another pitch a fifth and one octave higher could be produced on it by blowing more strongly (overblowing) as shown in notation I. This new pitch could be produced an octave lower by constructing a separate pipe two-thirds the size of the first one. If one then continued to construct pipes alternately four-thirds and two-thirds the length of the previous ones, an entire system of 12 notes could be generated, which is, with the exception of the means of creation, acoustically and proportionately in the same relation as is found in the Greek Pythagorean system. The English versions of the Chinese names for the 12 pitches seem quite fanciful; but they represent theoretically correct pitches, as do terms used in the Western traditional system, such as C or A-flat (Ab). The source of each name in the Chinese system is conjectural; but Chinese classical acousticians, like modern Western scientists, no doubt found value in creating a professional nomenclature that was divorced from everyday speech and potentially descriptive of the nature of the object. For example, the use of bell names may relate to the gradual preference for tuned bells over pipes in the music division of the courts. Names like "old purifier" and "equalizing rule" may refer to the pitch problems of the Pythagorean comma mentioned earlier.

A new interpretation of Chinese theory occurred in the late 20th century with the discovery of sets of 4th- and 5th-century tuned bells. Some of the bells produce two pitches and have the pitch names written at the two striking places. This information led to the development of a 12-pitch theory in which 5 pitches are generated in a cycle of fifths, and the 7 remaining pitches are located a major third above or below the first 4. If one starts from the Western C, the tones would appear as seen in notation II. The actual sounds produced on these ancient bells do not always match the pitch name given, but recent findings imply that it might have been possible to modulate to new pitch centres and different scales.

II

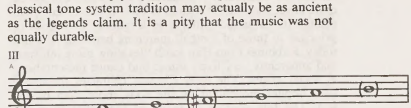


Scales and modes. For both Western and Chinese traditions, the 12 pitches are merely a tonal vocabulary from which a specific ordering of a limited number of pitches can be extracted and reproduced on different pitch levels. Such limited structures are called a scale. With a set scale it is possible to emphasize different notes in such a way that they seem to be the pitch centre. Such variations of pitch centre within a scale are called modes. In the Western traditional systems most scales use seven tones that can be transposed and that contain modes. For example, C major (C–D–E–F–G–A–B) can be made a Dorian mode by using D as the pitch centre without changing the pitches used (D–E–F–G–A–B–C), and the whole scale and its modes can be transposed to a higher or lower pitch level (F major, Eb major, etc.). The Chinese system concentrates in a similar way on a seven-tone scale but with a five-tone core (*wu sheng*) plus two changing (*pien*) tones, as shown in notation III.

Pitch centre and modes

The notes of a scale (a set of intervals not tied to specific pitches) are often indicated in Western music with syllables such as *do re mi*. The Chinese equivalent terms for notes in their classical scale are given in notation III. As in the Western system, modes can be constructed in Chinese music, and the scale can be transposed. From these comments it can be seen that the mythical emperor Huang-ti seems to have founded a very thorough system indeed. Throughout the Ch'in (221–206 BC) and Han (206 BC–AD 220) dynasties Imperial systems were tuned and retuned to meet Imperial and heavenly needs. As noted above, theoretical sophistications and experimentations continue on to the present day. How far back they may go in time is unknown, but in the late 20th century there have been discovered stone chimes from the 2nd millennium BC that imply by their tunings that the Chinese classical tone system tradition may actually be as ancient as the legends claim. It is a pity that the music was not equally durable.

III



Extramusical associations. Returning to the extramusical aspects of the Chinese system, one finds that the five fundamental tones are sometimes connected with the five directions or the five elements, while the 12 tones are connected by some writers with the months of the year, hours of the day, or phases of the moon. The 12 tones also can be found placed in two sets of 6 on Imperial panpipes (*pai-hsiao*) in keeping with the female-male (yin-yang) principle of Chinese metaphysics. Their placement is based on the generation of the pitches of each pipe

by its being either four-thirds larger or two-thirds smaller than the previous one, the smaller ones being female.

Classification of instruments. The Chinese talent for musical organization was by no means limited to pitches. Another important ancient system called the eight sounds (*pa yin*) was used to classify the many kinds of instruments used in Imperial orchestras. This system was based upon the material used in the construction of the instruments, the eight being stone, earth (pottery), bamboo, metal, skin, silk, wood, and gourd. The sonorous stones, ocarinas, and flutes mentioned earlier are examples from the first three categories. The bells are obvious metal examples. Another ancient member of the metal category is a large bronze drum (*t'ung-ku*), which is of special interest because of the widespread distribution of archaeological examples of it throughout Southeast Asia. Equally intriguing are the designs and sounds of the bronze head of the drum as well as the frequent statues of frogs around the rim of the head. Han dynasty military expeditions to the south report that bronze drums among southern peoples represented the spirit of rain and water and rumbled like bullfrogs. The possession of such bronze drums or later gongs was, and still is, prestigious among tribal groups in Southeast Asia.

Stringed instruments of ancient China belong to the silk class because their strings were never gut or metal but twisted silk. Drums are skin instruments, whereas percussive clappers are wood. One of the most enjoyable members of the wooden family is the *yü*, a model of a crouching tiger with a serrated ridge or set of slats along its back that were scratched by a bamboo whisk in a manner recalling the various scratched gourds of Latin American bands. The Chinese category of gourd is reserved for one of the most fascinating of the ancient instruments, the *sheng* mouth organ. Seventeen bamboo pipes are set in a gourd or sometimes in a wooden wind chest. Each pipe has a free metal reed at the end encased in the wind chest. Blowing through a mouth tube into the wind chest and closing a hole in a pipe with a finger will cause the reed to sound, and melodies or chord structures may be played. Many variants of this instrumental principle can be found in Southeast Asia, and it is not possible to know with assurance where this wind instrument first appeared. Western imitations of it are found in the reed organ and, later, in the harmonica and the accordion.

The mouth organ

Han dynasty: musical events and foreign influences. The extensive work in theory and classification in ancient times implies that there must have been an equally large amount of performance practice. Modern information on all these elements of music has suffered because of the destruction of many books and musical instruments under the order of Shih Huang-ti, emperor of the Ch'in dynasty. Yet there are several survivals from the Han dynasty that do give some insight into how the musical events took place. In the court and the Confucian temples there were two basic musical divisions: banquet music (*yen-yüeh*) and ritual music (*ya-yüeh*). Dances in the Confucian rituals were divided into military (*wu-wu*) and civil (*wen-wu*) forms. The ensembles of musicians and dancers could be quite large, and ancient listings of their content were often printed in formation patterns in a manner analogous in principle to those of football marching bands in America today. Rubbings from Han tomb tiles show more informal and apparently very lively music and dance presentations at social affairs. The early Chinese character for dance (*wu*) implies movement by the body more than by the feet. The folk sources of many of the songs from the *Shih Ching* and later books show that courtly musical life was not without its gayer and more personal and secular moments. The stringed instruments, notably the seven-stringed *ch'in* zither, apparently were popular as vehicles for solo music.

The Han dynasty empire expanded and at the same time built walls between its national core and western Asia. But these actions were paralleled by an increasing flow of foreign ideas and materials. Buddhism entered from India to China in the 1st century AD, whereas booty, goods, and ideas came from Central Asian Gandharan, Yüeh-chih, and Iranian cultures along the various desert trade routes via the cities of Khotan (Ho-t'ien) to the south (3rd through 5th century), Kucha (K'u-ch'e) in the centre (4th through 8th century), and Turfan (T'u-lu-p'an) to the north (5th through 9th century). Desert ruins and Buddhist caves from this period and later reveal a host of new musical ensembles and solo instruments. Two stringed instruments of particular interest are the angle harp (*k'ung-hu*) and the pear-shaped plucked lute (*p'i-p'a*). The harp can be traced back across Central Asia to the ancient bas-reliefs of Assyria. The lute also seems to have West Asian ancestors but is a more "contemporary" instrument. Variants of this instrument have continued to enter or be redesigned in China down to the present day. A delightful symbol of the long-term musical and commercial value of such a plucked lute is found on a 10th-century clay statue of a caravan Bactrian camel with two different styles of *p'i-p'a* tied to the saddle post on top of the rest of the cargo.

Diffusion of instruments in the Han dynasty

New percussion instruments are evident in the celestial orchestras seen in Buddhist iconography. One apparent accommodation between old Chinese and West Asian tradition is the *fang-hsiang*, a set of 16 iron slabs suspended in a wooden frame in the manner of the old sets of tuned stones. Knobless gongs related to the present-day Chinese *lo* seem to have entered the Chinese musical scene before the 6th century from South Asia, while the cymbals (*po*) may have come earlier from India via Central Asian groups. One of the most sonorous Buddhist additions was a bronze bell in the form of a basin (*ch'ing*) that, when placed rim up on a cushion and struck on the rim, produces a tone of amazing richness and duration. Among the varied new instruments pictured in heavenly ensembles, one can still find occasional "old-time" instruments such as a set of narrow wooden clappers (*ch'ung-tu*) tied together on one end like ancient wooden books. The clappers were sounded by compressing them quickly between the hands. Variants of this Chou dynasty instrument are still heard in all three major East Asian countries.

Not all the new influences in China came via religious or trade activities. During the Six Dynasties period (AD 220–589) China was rent by internal strife and border wars. The constant confrontations with the Tatars of the north caused an increased interest in the musical signals of the enemy via drums, trumpets, and double reeds. Although related instruments were equally evident to the south and west, there can be little doubt that the creation of cavalry bands with double kettledrums are direct imitations of the musical prowess of the horseback terrorists against whom the walls of China were built. With great effort and much blood, China gradually reunified under the Sui dynasty (581–618), and older courtly music and the latest musical fads were consolidated.

T'ANG DYNASTY

Thriving of foreign styles. The few centuries of T'ang dynasty existence (618–907) are supersaturated with brilliant Imperial growth and cultural flourishing as well as military and natural disasters. Such a rich loam of good and bad nourished one of the most fascinating eras of music history in the world. The more formal Imperial ceremonies revitalized the ancient orchestras of bells, stone chimes, flutes, drums, and zithers, plus large bands of courtly dancers. In reality, Imperial power was based perhaps less on the Mandate of Heaven than on the "liberation" of neighbouring countries, a development of more thorough tax systems and more and more trade cities and harbours. Into all these power sources flowed foreign goods and foreign ideas. Persians, Arabs, Indians, and Malayans were found in the foreign quarters of port towns, while every trade caravan brought in masses of new faces and modes of living. Perhaps it is not surprising that an 8th-century poet, Yüan Chen, should lament about air pollution created by western horsemen, about the ladies who studied western fashions and makeup, and about the entertainers who devoted themselves to only "western" music. (One must remember that the term western refers to the land west of the Great Wall.)

There was hardly a tavern in the capital of Ch'ang-an (now Sian, Shensi province) that could compete without the aid of a western dancing or singing girl with an ac-

Tavern entertainers

companying set of foreign musicians. Popular tunes of the period included "South India" and "Watching the Moon in Brahman Land," while beautiful, exotic dancing boys or girls were ever the rage. One set of girls from Sogdiana (centred in modern Uzbekistan) won the support of the emperor Hsüan-tsung (712–756) because they were costumed in crimson robes, green pants, and red deerskin boots and twirled on top of balls. Other girls from the area today called Tashkent inspired a poet of the 9th century, Po Chü-i, with their dance, which began with their emergence from artificial lotuses and ended with the pulling down of their blouses to show their shoulders, a style not unfamiliar to old Western burlesque connoisseurs. A study of the lithe bodies and flying sleeves on T'ang clay dancing figurines is an even more compelling proof of the style of the era. In such a context one can understand how eventually an additional character was added sometimes to the word for dance to indicate the movement of the legs as well as of the body.

In addition to all the commercial musical enterprises of the T'ang dynasty period, there was another equally extensive system under government supervision. The T'ang emperor Hsüan-tsung seemed particularly keen on music and took full advantage of the various musical "tributes" or "captives" sent to him by all the nations of Asia. This plethora of sounds was further enriched by the special area in Ch'ang-an called the Pear Garden (Li-yüan), in which hundreds of additional musicians and dancers were trained and in which the emperor himself was most active. Such trainees were often female. They followed in an earlier tradition of court girls (*kung-nü*) whose basic duties were to entertain distinguished guests.

The mass of different foreign musical styles in the capital was too much for the government musical bureaucracy. A distinction already had been made between court music (*ya-yüeh*) and common music (*su-yüeh*); but T'ang nomenclature added a third kind—foreign music (*hu-yüeh*). Eventually officials organized Imperial music into the 10 kinds of systems (*shih-pu chi*). Of these categories, one represented instrumentalists from Samarkand, whereas another group came from farther west in Bukhara (in modern Uzbekistan). Kashgar, at the mountain pass between the east and west, sent yet a different group. Musical ensembles also were presented to the emperor from the eastern Turkistan trade centres of Kucha and Turfan. India and two recently defeated kingdoms of Korea provided still other musicians. Chinese and Kucha music were blended by different musicians. One group was supposed to maintain the old styles of Chinese folk music, and there had to be one special group for the performance of formal Chinese court music. These 10 types by no means completed the picture, for nearly every Asian culture took its chance at musical goodwill in Ch'ang-an. Nothing from farther west appears in T'ang China, for culture hardly existed in Europe at that time. Nevertheless, one can sense in T'ang musical culture an internationalism not matched until the mass communications of the mid-20th century provided radio and phonograph owners with the delights of a similarly exotic and extensive choice.

Courtly music. The only music that can be discussed in a survey of a repertoire so large is the more official courtly music. Ritual presentations are generally divided into two types: so-called standing music, performed without strings and apparently in the courtyard; and sitting music, for a full ensemble played inside a palace. There are lists of the names of some pieces in these categories with their authorship usually credited to the emperor or empress of the time. For example, "The Battle Line Smashing Song" was said to be by the T'ang emperor T'ai-tsung (626–649). The accompanying dance is listed for 120 performers with spears and armour. A similarly grandiose piece is the "Music of Grand Victory" credited to the next T'ang emperor, Kao-tsung (649–683). Wu-hou (d. 705) is said to have written "The Imperial Birthday Music," in which the dancers form out the characters for "Long Live the Emperor" in the best modern marching-band tradition. Music inside the palace includes a concert version of "The Battle Line Smashing Song," with only four dancers, "A Banquet Song," and a piece supposedly composed by the empress Wu-hou in honour of her pet parrot, who frequently called out "Long live her majesty." Those familiar with music in the courts of Henry VIII and Louis XIV or with the songs always ending in praise of Queen Elizabeth I may recognize the cultural context of such music.

Courtly extravaganzas

Later-dynasty copies of T'ang paintings show ladies entertaining the emperor with ensembles of strings, winds, and percussion; and many of the choreographic plans of the larger pieces are also available in books. According to some sources, court orchestra pieces began with a prelude in free rhythm that set the mood and mode of the piece and introduced the instruments. This was followed by a slow section in a steady beat, and the piece ended in a faster tempo. Documents also tell much about the instrumentation and the colour and design of each costume of the musicians and dancers. No orchestral scores are to be found, however. One solo piece for *ch'in* survives, and 28 ritual melodies for *p'i-p'a* were discovered in the hidden library of the Buddhist caves of Tun-huang (Cave of the Thousand Buddhas), but the grand musical traditions of T'ang remain frustratingly elusive. Major clues to their actual sounds will be found in marginal survivals of such music, which will be discussed in the Korean and Japanese sections. The original traditions waned with the decline of T'ang good fortune, and the conflicts of the Five Dynasties and Ten Kingdoms period (907–960) brought the international period to an end.

SUNG AND YÜAN DYNASTIES

Consolidation of earlier trends. Despite the chaos of kingdoms in the 10th century, or perhaps because of it, cultural traditions solidified, so that by the Sung dynasty (960–1279) one can speak of a national rather than an international cultural mood. Many of the short-lived usurpers of regional governments were of "barbarian" (*i.e.*, Turkish) origin, but their general cultural efforts were to appear Chinese rather than to import further foreign fads. But one significant foreign musical addition of the period was from the northern Mongols in the form of a two-stringed fiddle, or bowed lute—the "foreign lute" (*hu-ch'in*). It became an important feature of the plebian theatre and teahouse world, which grew stronger and larger as more musicians and dancers were dropped from government payrolls. With the establishment of the Sung court, Confucian ceremonies and similar "old-fashioned" musical events were revived; but Imperial contributions to music of the period were primarily in the creation of gigantic historical or encyclopaedic works. For example, the official *Sung shih* (1345; "Sung History") contained 496 chapters, of which 17 deal directly with music, and musical events and people appear throughout the entire work. The *Yü-hai* encyclopaedia (*c.* 1267; "Sea of Jade") has 200 chapters, with 10 on music. It is interesting that the *lü* pipes are discussed separately under the topic of measurements. Manuals on how to play the seven-stringed *ch'in* zither also survive, as well as rare music collections such as the "Songs of Whitestone, the Taoist," based on the poems and songs of Chiang Kuei (1155–1221) and first printed in 1202. Many Sung poets continued to use the five- and seven-syllable-line *shih* form perfected by T'ang writers, which was believed to have been chanted to tunes strictly adhering to the word tones of the Chinese language. The singing girls (*chi-kuan*) of the teahouses and brothels and the general growth of urban, mercantile life inspired the creation of *tz'u* poems, which were free of word-tone restrictions, filled with colloquial phrases, and capable of freewheeling musical settings. A major source for music based on both the old and new forms is found in the rising world of public theatre.

Music theatre. Chinese drama can be noted as far back as the Chou dynasty, but it was really the T'ang period Pear Garden school that quite literally set the stage for Chinese opera. Regional music-drama flourished throughout the Sung empire, but the two major forms were the southern drama (*nan-ch'ü* or *nan-hsi*) and the northern drama (*tsa-chü* or *pei-ch'ü*). The *tz'u* poetical form was popular in both, although the southern style was held to be softer, with its emphasis on five-tone scales and flute and percussion accompaniments. The northern style

Northern and southern styles

is said to have preferred the seven-toned scale, to have used more strings, and in general to have been bolder in spirit. According to period writers, each of the four acts of a northern drama was set in a specific mode in which different tunes were used, interspersed with dialogue. The southern style was more lyrical.

The Mongols under Genghis Khan and later Kublai Khan finally succeeded in invading China, and the foreign Yüan dynasty (1206–1368) was founded. The two styles of drama noted above continued and intermixed under Yüan drama (*Yüan-ch'ü*), while the basic poetical form became *san-ch'ü,* popular songs of even freer style. On stage there appeared standard songs for specific situations or emotions that could be used in any opera, thus making it easier to communicate a story to mass audiences who may have spoken in many different dialects. Additional appeals to the general public were made by bringing onto the stage several forms of dancing and acrobatics, events that had been, along with several forms of puppet theatre, such gay parts of Chinese city life during the Sung dynasty.

MING AND CH'ING DYNASTIES

Internal Mongol struggles, natural disasters, and peasant revolts permitted the return of Chinese rule and the founding of the Ming dynasty (1368–1644). It in turn gave way to Manchu invasions from the north under which the last dynasty, the Ch'ing (1644–1911/12), was formed. Although there is much history and much blood involved in all such changes, one can view the music of these eras together under their two most active styles—theatre music and instrumental pieces.

Further development of opera. *Forms of the 16th–18th centuries.* The flourishing of regional music-drama has continued throughout the centuries from the Sung dynasty until the present day. Musically they vary greatly in their instrumentation and particularly in their voice qualities. However, all tend to follow a tradition of using either standard complete pieces (*lian-ch'ü*) or stereotyped melodic styles (*pan-ch'iang*) in every opera. The complete-piece approach of Yüan drama survives today primarily in a 16th-century form called *k'un-ch'ü.*

The *k'un-ch'ü* form

Nurtured in a more aristocratic form of theatre, the music of *k'un-ch'ü* was less bombastic than that of the popular theatre. The major instruments were the horizontal flute (*ti-tzu*) and the notched vertical flute (*hsiao*). The flutes often produce a special mottled tone by the presence of one hole that is covered by thin rice paper that buzzes quietly as one plays. The *sheng* mouth organ and the *p'i-p'a* plucked lute could also be found in *k'un-ch'ü,* along with a single free-reed pipe, *kuan.* The term *kuan* usually stands for one of several forms of double-reed woodwinds with cylindrical bore and no bell. Survivors of its ancient forms are found in Korean and Japanese court music. Variants of the single-reed *kuan* are found throughout Southeast Asia, where it is equally appreciated for its mellow, clarinet-like tone. A plebian instrument found in some *k'un-ch'ü* is the three-stringed plucked lute (*san-hsien*) with a snakeskin soundboard. Plucked with a bone pick, it enjoys great popularity in folk music as well as theatre music, and it developed in two sizes, the shorter one prevalent in the south and the longer one in the north. The shorter form is of particular historical interest, for it was imported into the Ryukyu Islands as the *jamisen* and from there moved to Japan, where it evolved into a samisen.

The vocal style of *k'un-ch'ü* matched the soft accompaniment and was usually performed by a male singing falsetto. Another style of opera from the same period, *i-yang ch'iang,* seemed more appealing to the general public and is noteworthy for its use at some point in its development of a chorus (*pang-ch'iang*) as well as of soloists. In addition, passages in colloquial speech were often interpolated between lines of classical poetry in order to explain them. Such lines were often sung. Still another Ming music-drama genre of considerable influence in the myriad regional forms is the clapper opera (*pang-tzu ch'iang*). In addition to the rhythmic importance of the clappers, the instrumental accompaniment of this form is noted for its emphasis on strings, the principal form being the moon guitar (*yüeh-ch'in*), a plucked lute with a large, round wooden body and four strings in double courses. An interesting addition to this instrument is the presence of a thin strip of metal tied at both ends inside the body to give the instrument a richer tone. Among the endless variants of style and accompaniments in Chinese regional opera, one must add the sounds of the extremely large flat gongs heard in the southwest and the *yang-ch'in* (western zither), particularly popular in Cantonese music. The latter is often called a butterfly harp, though it is neither a harp nor a butterfly but a hammered dulcimer derived from a Middle Eastern instrument (*santūr*) brought into China in the 18th century. Each of the myriad types of regional opera flourishing in China developed vocal styles and orchestrations that helped make it distinctive. With informed practice, listeners can still distinguish regional vocal styles, which vary from low, sensual sounds to high and nasal falsettos. For the rest of the materials concerning theatre music, it is best to turn to the primary music-drama form since the 18th century, Peking opera (*ching-hsi* or *ching-chü*).

The clapper opera

Peking opera. Credit for the beginning of Peking opera is given to actors from Anhwei appearing in Peking in the 1790s. However, Peking opera really combines elements from many different earlier forms and, like Western grand opera, can be considered to be a 19th-century product. In addition to all the instruments mentioned above, many others may be found.

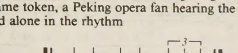
The most common melodic instrument for opera is some form of fiddle, or bowed lute (*hu-ch'in*). It comes in several different forms, such as the *ching-hu* and *erh-hu.* Although the shape of the body may be different, all traditional Chinese fiddles exhibit certain specific structural characteristics. The small body has a skin or wooden soundboard and an open back. The two strings pass over a bridge and then are suspended above a pole to the pegs, which are inserted from the rear of the scroll (not from the sides as on a Western violin). Such a system places one string above the other rather than parallel to it (as on a banjo or a *p'i-p'a*). Because of this, the bow passes between the strings, playing one string by pressing down and the other by pulling up. The fingerings of tunes are done by sideways pressure, along the strings; they are too far from the pole for it to serve as a fingerboard, which, because of the vertical stringing, would be a nuisance in any case. It is this unique manner of fiddle construction that helps one determine the source of many of the bowed lutes of Southeast Asia.

Barrel drums with tacked heads (*ku*) and a double reed with a conical bore and bell (*so-na*) are used in military scenes, along with cymbals (*po*) and large flat gongs. The most common percussion instruments are a small flat gong (*lo*), a drum (*pan-ku*), and clappers (*p'ai-pan*). The small gong is some eight inches in diameter; the face is slightly curved except for a flat centre spot. It is designed in this manner in order that the tone and pitch of the gong will rise quickly each time it is hit. This "sliding" gong effect is characteristic of the Peking sound. The *pan-ku* or *tan-pi ku* is equally unique in construction. The skin is stretched over a set of wooden wedges strapped in a circle with only a small spot in the middle completely hollow. This allows the performer to produce a very dry, sharp sound. Such a tone is practical as well as aesthetic, for the *pan-ku* player is often the leader of the ensemble, and his signals are essential to the coordination of the performance. The drum player frequently plays the clapper as well, holding the clapper in his left hand while playing the drum with a narrow bamboo stick held in his right hand.

The role of rhythm

In all East Asian music one must remember that harmony and harmonic progression are not parts of traditional music. The functions of harmony—such as underlining expression, providing sonic contrast, and creating a sense of forward motion—are handled with equal efficiency by rhythm in East Asia, although the methods and sounds are very different. In both traditions, the choices are not arbitrary, and with cultural exposure one comes to recognize the musical intention, even though it is not necessary to know precisely what chord or what rhythm pattern produces an appropriate musical effect. For example, very

few listeners to Western music know that a doubly diminished chord (C–Eb–Gb–A) played tremolo means danger, although all would recognize the danger signal by ear. By the same token, a Peking opera fan hearing the large gong played alone in the rhythm



would know that the situation is a similar moment of confusion but probably would not know that the pattern is named the scattering hammer (*luan-chüeh*). Pattern names are for specialists, but pattern sounds and "meanings" are for attuned listeners. Other aspects of the functions of rhythm in East Asia will emerge in the examination of other cultures. For the moment, attention will be given to the melodic side of Peking opera.

Like any theatrical music, the tunes of Peking opera must conform to the text structure and the dramatic situation. In the latter case, one finds that a majority of Peking aria texts are based on series of couplets of 7 or 10 syllables each. Although there may be several verses set in strophic form (*i.e.*, music repeated for each strophe, or stanza), part of the musical tension is maintained by the interjection of comments or short dialogue between the two lines of each verse. These leave the listener waiting for the completion of the line. The tune aids in this forward motion and tension by playing what could be considered an incomplete melodic cadence (point of resolution) at the end of line one, which is brought to a final resolution at the end of the second line. From a dramaturgical standpoint, the arias of Peking opera can be categorized into different types whose style is recognizable in the same way that one can tell, without language ability, the mood of a love, farewell, or vengeance aria in Italian opera.

Prototypes of Peking melodies

Peking melodies themselves tend to fall into two prototypes called *hsi-p'i* and *erh-huang*. Within each of these general types there are several well-known tunes, but the word "prototype" has been used to define them, as each opera and each situation is capable of varying the basic melody greatly. The two basic identifying factors are the mode of the melody and the rhythmic style of the accompanying percussion section. In general, serious and lyrical texts are performed to an *erh-huang* melody and *hsi-p'i* tunes appear in brighter moments, though in such a large genre there are many other possibilities. Notation IV contains the string introductions to examples in the two basic types. They are transposed to the pitches of notation II for the sake of comparison. In actual performance the fiddle may be tuned lower for *erh-huang* melodies. How do the tunes differ? Both emphasize the pentatonic core and have a "changing" tone B (its pitch is actually between the Western B and B♭), but their modes differ. *Hsi-p'i* are said to emphasize (in the context of our transposition) E and A, and *erh-huang* G and C. *Hsi-p'i* melodies are often more disjunct. Although both examples are set at a standard tempo (*yüan-pan*), the *erh-huang* is faster and its rhythm denser, as it is a male aria, while the *hsi-p'i* is female and slower.

Both pieces could be played at a slower (*man-pan*) or faster (*k'uai-pan*) tempo, however, or could be accompanied by other special rhythms. Such choices often cause changes in the melody itself. In general, the choice of both tune and rhythm style is guided by the text and the character. In most arias each sentence is separated by an instrumental interlude.

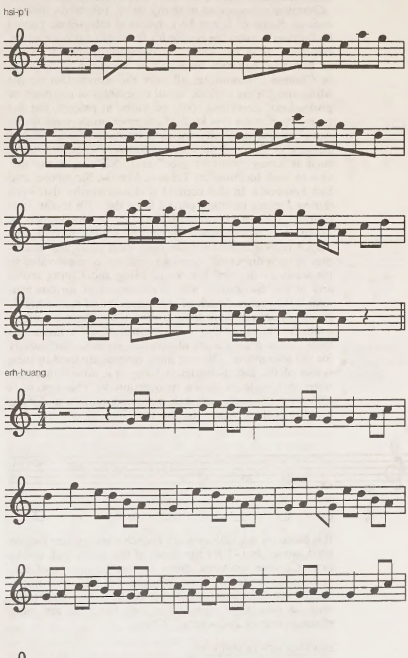
Peking opera is also characterized by colourful costumes and striking character-identifying makeup as well as acrobatic combats and dances. These conventions of Chinese opera are similar to those of 18th-century European traditions, though the sounds are certainly quite different. The need to communicate in music or in theatre requires the repeated use of aural and visual conventions if an audience is to understand and be moved by the event.

Other vocal and instrumental genres. The emphasis here has been on opera because it is best known, but there are many other popular forms from the Ming and Ch'ing periods. One is storytelling (*shuo-shu*). This tradition, which is as old as humankind and is noted in China's earliest books, continues in China in a purely narrative form, in a sung style, and in a mixture of the two. Until the advent of television and government arts control, there were narrators who recounted traditional stories in nightly or weekly segments. Their idiom was like that of surviving tellers of shorter stories. The text is usually in rhyme and is spoken in rhythm. Chinese storytellers may perform unaccompanied, but generally at least a clapper rhythm is present. One string instrument, such as a three-string *san-hsien* or four-string *p'i-p'a* lute, is also common. Songs accompanied by a drum (*ta-ku*) are the best known. The narrator not only relates the story but usually plays the clappers and a drum as well. Since the text is the core of the genre, standard melodies are used. Additional accompaniment may be provided by a string ensemble like that of opera.

Musically, the various shadow- and hand-puppet plays also are similar to the opera tradition except that, as in Southeast Asian puppetry, a manipulator must often be the singer-narrator as well.

These genres, like many regional opera forms, are often performed on temporary street stages and are eclectically creative. Saxophones and other Western instruments may combine with the ubiquitous Chinese fiddles and percussion instruments. Topical popular tunes and well-known Western music can appear among opera melodies as the drama unfolds. Recordings mix with live music so that, for

IV



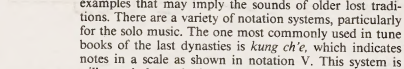
example, a battle scene may be accompanied by Chinese percussion sounds, firecrackers, and a recording of Nikolay Rimsky-Korsakov's "The Flight of the Bumble Bee."

Leaving the many forms of vocal and theatrical music, it is appropriate to turn briefly to the instrumental. The 25-stringed *se* zither, with movable bridges, and the seven-stringed *ch'in*, with permanent upper and lower bridges (like a piano), were well known for solo music in ancient times. During the last dynasty, collections of *ch'in* music and instruction books flourished as part of certain neo-Confucian revivals. Many musical notations were developed, perhaps the most interesting variety for the *ch'in* being one in which Chinese characters were artificially constructed by combining symbols for the notes with indications of fingering technique, such as up strokes, down strokes, or harmonics. Although most of the music was based on vocal pieces or evoked some scene, there were several examples of variation forms that had an important influence on Korean and Japanese forms that followed. The *p'i-p'a* likewise developed an extensive repertoire of solo pieces, many of them quite virtuosic and pictorial. For example, anyone hearing a *p'i-p'a* battle piece needs to know very little Chinese to recognize the musical interpretations of the action. Since the mid-20th century there has been a considerable revival of solo literature for the *cheng*, a zither with 16 strings and movable bridges whose popularity spreads as far south as Vietnam. The strings are apparently influenced by the Middle Eastern dulcimer mentioned above (*yang-ch'in*), for they are metal.

Music for the *ch'in* zither

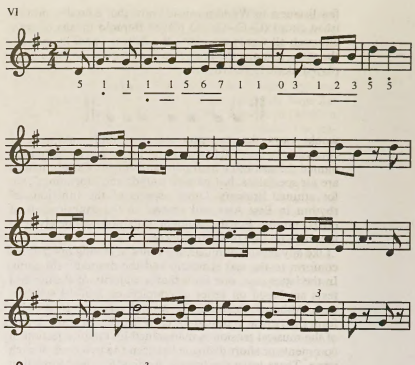
Chamber music exists in many styles, functions, and locations. Some of it can be considered folk music played by farmers or working people for festivals or private entertainment, as in the American bluegrass tradition. Music of this type can still be heard at weddings or funerals in Chinese communities all over the world. During the Ming and Ch'ing periods, small ensembles of courtiers or professional musicians could be found at palaces, but the major sources for this kind of chamber music were in the world of the musically inclined businessman or trader. Because of this, certain regional forms of chamber music such as Amoy "southern pipe" and Shantung music survive in such locations as Taiwan, Manila, Singapore, and San Francisco. In this context it is noteworthy that even during Japan's isolation period from the 17th to the 19th century, Chinese vocal and chamber music, known in Japanese as *minshingaku* (Ming and Ch'ing music), was played in Nagasaki, the only open port in Japan. Examples of such dispersed regional music are of great value in the study of the oral history of Ming and Ch'ing music and of the distribution and development of various musical instruments. Much of the repertoire of such stylistic groups is derived from theatre music, but there are many examples that may imply the sounds of older lost traditions. There are a variety of notation systems, particularly for the solo music. The one most commonly used in tune books of the last dynasties is *kung ch'e*, which indicates notes in a scale as shown in notation V. This system is still popularly used, although mainland sources prefer the number system shown in the first line of notation VI.

V



It is based on the 19th-century French *chevé* system (which used numerals 1–7 for the notes of the scale) and, unlike other Chinese notations, shows rhythm by the use of dots and beams borrowed from Western 8th and 16th notes. Percussion accompaniments also can be found in a similar style, as can larger ensemble scores, but both are more characteristic of 20th-century China.

VI



DEVELOPMENTS SINCE 1911

Period of the Republic of China and the Sino-Japanese War. Under the influence of missionary and modernization movements, many musical experimentations occurred in the last dynasty, but these were greatly increased by the rise of the first republic in 1911 and the establishment of communist rule in 1949. During the period of the republic and of the Japanese war, a plethora of new songs were created in "modern" style, the most famous being shown in notation VI. The piece, "March of the Volunteers," was written in 1934 by Nie Er (Nieh Erh) to text by the modern Chinese playwright Tian Han (T'ien Han) as a patriotic march and was adopted as the national anthem in 1949. It is an excellent example of a mixture of new and traditional Chinese music. The first phrase implies a major mode with its use of F♯. However, after that point the entire piece is Chinese pentatonic. The first phrase also leads one to expect symmetrical four-bar phrases, but the tune quickly takes a more flexible Chinese course. Chinese and Western composers continued to try out bits of each other's traditions with only occasional success, and individual Chinese artists have become famous for their performance on Western instruments. Chinese instruments in turn have been subjected to many modernizations, such as the building of a family of *erh-hu* fiddles by the creation of bass and alto versions. In conjunction with this movement there was the appearance of concerti for such instruments accompanied by a mixed Western and Chinese orchestra.

The "modern" style

Communist period. As was noted earlier, many completely traditional forms continued, particularly in foreign Chinese communities. The special point of interest since 1949, however, is the application of Marxist doctrine to the musical scene of China. The first obvious area of change is found in the ever popular forms of regional and Peking opera. Although the appeal of traditional tales of emperors, princesses, or mythological characters could not be suppressed, the emphasis of all new operas was on workers, peasants, soldiers, and socialism. Thus, *San-kuo chih yen-i* (*Romance of the Three Kingdoms*) or *K'ung-ch'eng chi* (*The Ruse of an Empty City*) tend to be replaced by *Qixi Baihutuan* (*Ch'i-hsi Pai-hu t'uan; Raid on the White Tiger Regiment*) or *Honghu chiwei dui* (*Hung Hu ch'ih-wei t'ui; Red Guards of Hung Lake*). Aria topics also vary, such as "Looking Forward to the Liberation of the Working People of the World" or "Socialism Is Good."

As part of the encouragement of people's music, the national government emphasized regional folk music. Provincial and national research institutes were created to collect and study such music, and folk songs were incorporated into primary as well as advanced and Western music education. In general, folk music was "reconstructed" away from its former individualistic nature into collectives

of choruses or folk orchestras. The topics of such regional songs also were reconstructed so that they reflected the new socialist life. The most famous new folk song from Shensi province is "Red Is the East," while the Miao people were credited with "Sing in Praise of Chairman Mao." During the Maoist period, more than 50 minority groups and provincial Chinese ensembles had at least one song directly in praise of Chairman Mao, while other songs dealt with local industries and accomplishments. Such songs are sometimes performed in regional style with traditional accompaniments, although they may often be found arranged Western-style for use in the public schools of the nation. This effort, in addition to the number of recordings that are available, make it possible for a Chinese citizen to become aware, perhaps for the first time in history, of the great variety of local music traditions in his large country, even though such music appears now in Marxist reconstructions. Marxist defense of this changed folk music is that music of a given period must reflect the views and aspirations of the masses (as understood by the government) and must be based on idioms of the people. Composers of concert music have produced many folk orchestra compositions along with symphony, piano, and military band music based on this basic Marxist musical principle, called Socialist Realism. When dealing with traditional instruments and vocal styles, the composers have sometimes created extremely original and interesting pieces despite the general conservatism of government aesthetics policies. Vocal and choral music are preferred because of their ability to communicate specific national goals more efficiently than, for example, *The Sacred War Symphony.*

Socialist Realism

It must be remembered that music exists in a cultural context and that it has never remained static since the world began. In the late 20th century, music of all periods from every society is available to those with sufficient mass communication sources. Exchanges have been made between Western and Chinese ensembles and musicians, and audiocassettes and radio broadcasts cannot be easily silenced. Euro-American music is part of China's urban culture, and new socialist messages can be heard in Western-style popular music settings. At the same time, tentative efforts have been made to use contemporary Western idioms in Chinese concert music. It does seem unlikely that the tuning of the *lü* pipes for rulers will ever be a major concern of Chinese musicians again, but the ability of China to preserve so many historical facts, materials, and idioms along with modern changes is sufficient to keep the musical world in awe for some centuries to come.

The music of Korea

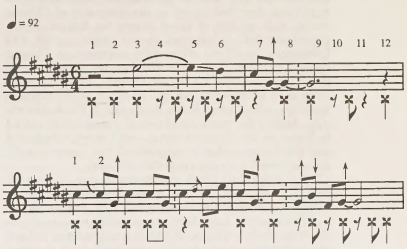
On a map Korea looks like a finger pointing from the top of China down to the lower part of Japan. Thus, one would expect its music to reflect its "bridge" position between two such powerful traditions. The movements of foreign, particularly Chinese, armies and cultures are indeed major factors in Korea's tradition. But beneath these reflections lies a deeper core of indigenous musical styles that, at first hearing, seem most strange to the ear of listeners with preconceived notions as to what East Asian music sounds like. A possible additional factor in the growth of Korean music is the country's position as a peninsula jutting out from Manchuria and from the spawning ground of many Mongolian hordes. Archaeological sources indicate that various Mongol peoples from northern Asia did indeed occupy areas of Korea from at least 2000 BC, and Chinese writings show that their people and armies were active in Korea from the period of the Chinese Han dynasty (206 BC–AD 220) on. Obviously, a study of Korean music contains riches extending far beyond its geographic borders.

SHAMAN MUSIC

The earliest references to music in Korea are found in a 3rd-century-AD Chinese history book that comments about agricultural festivals (*nong-ak*) with singing and dancing among the tribes of northwestern Korea. Such events are still a strong part of Korean life. Another ancient but long-lived tradition in Korea is shamanism, or communication with the unseen world by medicine men in a state of trance. This is of special interest because such a belief is characteristic not only of all northern Asian tribes but also of other peoples (such as Eskimos) who live in the northernmost regions of the world. Korea is one of the few countries south of the Arctic area that maintain strong shamanism in the face of several foreign religious adoptions such as Buddhism, Confucianism, and Christianity. A few Korean shamanistic events, however, show still other possible prehistoric connections. For example, sometimes the head of an animal may be placed in the centre of a Korean ritual ground, and mythology says that the first ruler of Korea was created by the union of god and a bear. Divine origin is part of Japan's Imperial mythology, and the head of a bear is central to the important rites of the Ainu tribal culture in northern Japan.

Today, a female Korean shaman (*mudang*) may use many combinations of musical instruments. The simplest and potentially most significant accompaniment is a small, flat gong with a slight rim. It brings to mind the single-headed pan drum with a wooden or bone hoop found in the shamanism of most of Central Asia and in the Arctic Circle as far away as Lapland and Hudson Bay. A drum sound itself is produced in Korea by the most popular percussion instrument, the *changgo,* an hourglass-shaped, two-headed drum struck by the hand on the left head and a stick on the other. In Korean shaman rituals flutes, double reeds, fiddles, and other gongs and drums may be used that at first sight may appear rather Chinese. The sound, however, creates a totally different impression.

VII



The music shown in notation VII represents a flute and gong excerpt from a shaman ensemble that also contains a drum, zither, fiddle, and oboe in a driving polyphony (combination of simultaneous voices, or parts) that seems closer to Dixieland jazz than to Chinese music. The flute tune, with its microtonal slides, its use of a fourth (c♯″-g♯″) "out of tune" with the Chinese *lü* pipes, and its rhythm, produce a very un-Chinese, jazzlike sound. The style is totally traditional and is notated to show the special Korean proclivity for a 6-beat unit, which in this excerpt is particularly strong in the accompaniment of the other instruments. The dotted bar lines in the transcription imply a kind of polymetric syncopation that often gives Korean folk and popular music its special appeal. One part seems to be in 4 whereas others are in 6, so that they come together only after 12 beats. Triplets and even 5-beat forms are found as well. Thus, it is evident from a brief look at one example from one type of Korean shaman music, in addition to a discussion of other forms and characteristics, that Korean cultural materials continue to reflect fascinating mixtures and mysteries. The total style cannot be called purely Chinese or northern Asian but simply Korean. The characteristics of the first example are typical of the kind of music best known and loved by the general Korean populace. Studies in Korean music, however, have tended to concentrate on the less familiar styles of court music, which are maintained by dedicated national music institutes and Korean scholarship.

Polyphony, tuning, and rhythm

COURT INSTRUMENTAL MUSIC

According to legend the Three Kingdoms of Koguryŏ in the north, Paekche in the southwest, and Silla in the southeast were established in the century before AD 1 along with a small Japanese-related enclave (Kaya; Japanese: Mimana). The subsequent organization of courts and the introduction of Chinese religions resulted in an ever-increasing importation of Chinese musical materials. Indications of this can be found in such sources as paintings in an AD 357 tomb near An-ag, a colony of China at that time. A horseback band of the Chinese Han dynasty style is seen with drums and a small bell hit with a hammer. One brave rider apparently is able to play the Chinese panpipes in transit. Deeper in the tomb a zither, a lute, and a very long end-blown flute can be seen accompanying a dancer whose long nose and costume imply that Central Asian traditions may have traveled even as far as Korea by the 4th century. The Silla dynasty domination (668–935) coincided closely with the heyday of the Chinese T'ang period, and the subsequent Koryŏ (935–1392) and Chosŏn (Yi) dynasties (1392–1910) also tended to match parallel Chinese periods. Thus, Korea's court-music traditions tended to reflect those of China.

Survivals of ancient Chinese music

Being on the border of Chinese culture, Korea was able to maintain certain ancient traditions during periods of barbarian domination in China proper. Such marginal survivals are of particular importance because many have continued to the present day, thus giving extremely rare examples of music traditions long gone from the land of their origin. For example, in the Silla period, court music was divided into *hyang-ak,* Korean music; *t'ang-ak,* T'ang and Sung Chinese music; and *a-ak,* Confucian ritual music. The instruments used for these ensembles were those described earlier in the discussion of Chinese music, such as sets of tuned stones (in Korean *p'yŏngyŏng*) and bells (*p'yŏnjong*), mouth organ (*saeng*), and instruments in all the other eight categories of Chinese classical traditions (*e.g.,* those based on the materials used in their construction; see above). Unlike China, ensembles using such instruments can still be heard in the national institutes of North and South Korea and in Confucian rituals. Among the many instrumental treasures still played in Korea is the *ajaeng,* a zither—with seven strings and movable bridges—that is not plucked but, more remarkably, is bowed with a rosined stick of wood.

The globular flute (*hun*), mentioned as one of the very earliest artifacts of Chinese music, has been played in Korean Confucian temples since the 12th century, as has a *chi* flute, which has a bamboo mouthpiece plugged into the mouth-hole with wax. In addition to five finger holes it has a cross-shaped hole in what on other flutes is the open lower end. The lower end of the *chi* can thus be closed by the little finger of the left hand. This unique flute is known to have been in Korea by at least the 11th century and, like the previous two examples, has totally disappeared from the rest of East Asia. By contrast, the *p'iri* cylindrical double reed aerophone (wind instrument) has many relatives in Asia, but the rich saxophone-like tone produced by its deceptively narrow tube body and large reed are not heard elsewhere. It is heard in many other forms of Korean music, from folk festivals to party music.

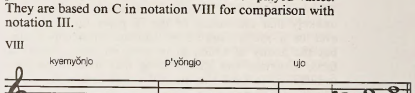
Indigenous instruments

Not all the instruments of Korea are Chinese imports. The *taegŭm* flute with six finger holes, a membrane-covered hole for a buzzing sound, and open holes near the end would seem to be a Chinese instrument, except for its spectacular length of 74 centimetres (2 feet 5 inches) and its gigantic mouth-hole. A Korean musician, Wang San-ak, is credited with the invention in the 7th century of a *kŏmungo* zither with six strings. Two strings on one side and one on the other have movable bridges, whereas the central three strings pass over 16 bridges. It is played by plucking the strings with a wooden stick. One of the other mysterious Korean instruments is the *haegŭm* two-stringed fiddle. It would seem to be an obvious relative of the Chinese equivalent, with the bow passing between the strings, except that the neck is bent toward the strings (rather than away from them, as in the rest of the world), and the pegs seem to be inserted backward, so that the strings are wrapped around the large round part of the pegs instead of the narrower end, which sticks out, unused, from the back of the neck. The *kayagŭm* board zither with 12 strings and movable bridges is surprising in sound to those accustomed to Chinese and Japanese zither melodies. It is held that the instrument was created in the 6th century in the Japanese-dominated Kaya area—thus the survival of one example in the 8th-century Japanese Shōsō-in treasure house. The *kayagŭm* is regarded as Korea's favourite native instrument, and it can be heard in all levels of Korean music and dance. The *sanjo* variation forms for this zither represent one of Korea's most famous purely instrumental genres.

Much of what is known about the origin of instruments is derived from Chinese and Korean historical books and administrative documents, such as the grand list of presents sent by the Chinese Sung emperor to Korea in the year 1111. The list includes 10 sets of stone chimes and 10 bell sets, along with 5 iron equivalents and numerous other instruments. Korean musicians performed successfully at the Chinese court, and Korean monks attended the international training centres in China to learn Buddhist chant. During the reign of Sejong (1419–50), new Imperial shrine music and traditional Confucian ritual music were emphasized along with musical settings of epics written in the newly developed Korean alphabet. The grand traditions of China were preserved under the guidance of the court master of music, Pak Yŏn (1378–1458).

Proof of the diligence and concern of these and later efforts are found in the *Akhak kwebon* ("Music Handbook"), first appearing in 1493. The nine volumes of this work contain pictures of all the court instruments along with their fingerings or tunings, costumes and accessories for ritual dances, and the arrangements of dance designs and orchestral seatings. The first three volumes deal with music theory and contain ample evidence of the continuation of the complete Chinese classical tradition discussed earlier. The only differences are the pronunciations given to the Chinese characters in which the terminology is written. Over the centuries the naming and interpretation of the pentatonic scales in Korea have varied greatly. Today *kyemyŏnjo* and *p'yŏngjo* are considered basic. *Ujo* is a variant on *p'yŏngjo,* usually a fourth higher. The exact pitch on which these modes are written or played varies. They are based on C in notation VIII for comparison with notation III.

VIII



Balance of Korean and Chinese traditions

From the short theoretical discussion above it should be evident that Korean musicians maintained a balance between native and Chinese traditions. Such a balance is seen in the standard instrumentation of the three major court orchestras. By the 15th century the Chinese-style *t'ang-ak* and the Confucian *a-ak* ensembles concentrated on Chinese instruments such as bell and stone chime sets, and the texts of surviving *t'ang-ak* pieces such as *Loyang-chun* ("Spring in Loyang") follow the Chinese *tz'u* poetical form. Processional military music (*chui-ta*) begins in the style seen in ancient drawings, with drums, gongs, and accompanying conch shell and straight trumpets, in addition to a "barbarian" oboe with a conical body. This ensemble is followed by a softer one with the more typical Korean hourglass drum (*changgo*) and cylindrical oboe (*p'iri*), the unusual Korean fiddle (*haegŭm*), and flutes (*taegŭm*). The softer ensemble can also be heard in dance music (*samhyŏn*), whereas chamber music (*chŏng-ak*) softens this group further by using a smaller oboe along with the later addition of the Chinese "western" dulcimer (*yang ch'in,* or, in Korean, *yang-gŭm*). The most famous suite of movements in this and in orchestral traditions is the *Yŏngsan hoesang,* which consists of 9 to 11 pieces taking some 30 minutes to play. The title is based on a former religious chant about the Buddha preaching on Mount Yŏngsan, but the pieces attached to this general name have since lost their vocal tradition. When the Korean ensem-

ble (*hyang-ak*) plays pieces like *Sujech'ŏn* ("Long Life as Immeasurable as the Sky"), one hears a more indigenous combination of the hourglass drum, oboes, flutes, fiddles, and the special bowed zithers (*ajaeng*). Although the style is still slow and "ancient," the sounds seem less Chinese.

The survival of so many old traditions is partly due to the preservation of notation books. Many are in the traditional Chinese forms. In the late 15th century, however, a Korean mensural system (*i.e.*, a notation showing time values) was created that, through the use of columns of 16 squares, gave a clearer indication of rhythm and tempos than do most Chinese notations. This system, usually with modifications into 6- or 12-square groups, is used today to notate the six-beat rhythms and can be read as easily as Western notation if one can read Korean.

VOCAL MUSIC

Vocal music is another important side of the Korean tradition. One of the longest and rarest older forms is the *kagok*, which consists of 26 five-line solo songs and one duet. Accompaniments and interludes are provided by a small ensemble. *Sijo* is a classical three-line form of Korean poetry that also can be sung to the accompaniment of the hourglass drum. Narrative songs are found in the genre called *kasa* accompanied by a flute and drum.

Kagok, sijo, and *kasa* are all types of court music. The dominant narrative form of music performed today, however, is the folk genre *p'ansori*. It is traditionally performed by a singer-narrator (*kwangdae*) and a drummer (*kosu*), who marks phrases with rhythmic patterns on a barrel drum (*puk*) and with vocal interjections (*chuimsae*). Spoken narration and dialogue (*aniri*) are balanced with songs (*chang*), body movements, and fan gestures as hours of epic drama unfold. It may be that *p'ansori* emerged from ancient shamanistic entertainments of the gods before it became popular with the aristocracy. The earliest written records of it date to 1775. Of the many stories noted in later sources, five have survived both in written form and in popular folk tales. In the late 20th century, government support revived the tradition, so that *p'ansori* epics are available on recordings and in professional performances.

P'ansori

MODERN MUSIC

During the period of Japanese occupation (1910–45), indigenous arts were suppressed and Western music dominated the education system. As in Japan, this dominance continues, and Korean skills in Western music performance have earned international recognition. Since World War II, the support of national arts in Korea has created many new tradition-based musics. Those in North Korea follow the Chinese pattern discussed earlier, while the South generally follows Western contemporary trends.

The music of Japan

As has been noted, Japanese music can be considered a national tradition set in the satellite category of the general East Asian music culture. Korea served as a bridge to Japan for many Chinese musical ideas as well as exerting influence through its own forms of court music. A comment has been made as well about the presence of northern Asian tribal traditions in Japan in the form of Ainu culture surviving on Hokkaido island. It should further be pointed out that the island isolation of Japan allowed it to develop its own special characteristics without the intense influences of the Chinese giant and the Mongols so evident in other mainland cultures. Therefore, in the ensuing discussion all the "foreign" elements will be placed in the matrix of traditions and styles that are characteristically Japanese.

MUSIC BEFORE AND THROUGH THE NARA PERIOD

Early evidence. As in the case of the history of mainland traditions, ancient Chinese sources and modern archaeological data provide the earliest surviving insights into Japanese music. Archaeologists have discovered materials of Neolithic people in Japan and pottery remains of the so-called Jōmon culture dating back as far as the 5th millennium BC. Among the items recovered from the subsequent Yayoi period (3rd century BC–3rd century AD), the musically most significant finds are *dōtaku* bronze bells. They show that the native population had adopted Chinese metallurgy. The shape of the bells and the locations of their remains indicate that they may have entered the Japanese islands with tribes migrating from northern Asia.

Archaeological evidence

The gradual domination in Japan of one group called the Yamato clan became more evident in the Tumulus period (*c.* AD 250–*c.* 500) and led to the present Imperial system. Specific evidence of its musical life is found first in certain tomb figurines (*haniwa*), which were substitutes for the earlier Asian tradition of human sacrifices at the death of a leader. One *haniwa* has been found playing a barrel drum with a stick, while another figure is seated with a four- or five-stringed board zither across his lap. Crotal bells (pellet or jingle bells) are found on costumes, and some statues seem to be of singers. The zither is of special interest, for it is related to the Korean *kayagŭm* mentioned earlier as appearing in the Japanese section of Korea (Kaya) by at least the 6th century. It also may be the earliest example of the *wagon,* or *Yamato-goto,* a six-stringed zither with movable bridges found in Japanese Shintō music. The crotal bells survive in the form of the *suzu* bell tree, an instrument characteristic of Shintō dances only. The interpretation of another figure as a singer and the presence of a drummer are rather too general for conclusions, although a Chinese history book of the 3rd century (*Wei chih,* AD 297) does speak of the natives of Japan as singing and dancing during a funeral. This source also notes two actions well-known in Shintō today: a concern for purification and the use of hand claps in praying before a shrine.

The mention of shamanism also is found in Chinese accounts and is of particular interest to those concerned with the northern Asian aspects of Japanese culture. In this context it must be remembered that the Ainu were as populous and strong as the new Japanese people at the time of the founding of the Yamato dynasty. Battles between the Japanese and Ainu are noted in 6th-century Chinese books such as the *Sung shu* (513) and, rather like 19th-century American Indians, Ainu were found as mercenary troops in a group of Japanese forces sent to assist the Korean Silla kingdom in the 7th century. The Chinese *Sui shu* history book (630) mentions tattooed people like the Ainu, as well as a five-stringed zither and a flute. Ainu culture today maintains a Jew's harp and no flutes, but it does have a *tonkori* zither with two to five strings. It is unlike the zither on the lap of the earlier tomb figure in both its shape and playing position, being held like a banjo and played open-stringed with both hands. The surviving shamanism of the Ainu has equivalent forms in early Shintō and in a few surviving Japanese folk "mountain women" traditions. However, the guttural vocal style and the frequent polyphonic textures of modern Ainu music today seem culturally to point north rather than south or west. Perhaps the Ainu are a living link between present-day civilization and the life pictured in ancient Chinese documents.

As the Japanese people gradually drove the Ainu northward, they solidified their own internal structure and established stronger ties with continental culture. Records show that a Korean Silla (in Japanese, Shiragi) emperor sent 80 musicians to the funeral of a Japanese ruler in 453. Chinese Buddhism was officially introduced as a religion in Japan in the 6th century, selected converts being sent to China for proper training in the rituals (hence the music) of that faith. In 612 a Korean musician, Mimaji (in Japanese, Mimashi), is believed to have introduced masked dances and entertainments (*gigaku*) and southern Chinese music (*kuregaku*) into the Japanese court. Finally, by the 8th century Japan produced its own first written chronicles, the *Koji-ki* (713; "Record of Ancient Matters") and the *Nihon shoki* (720; "National History"), which recount the mythological origin of music as the form of an entertainment used by the gods to tempt the sun goddess out of her hiding in a cave (see also JAPANESE LITERATURE). Indirect references to music appear in semi-historical accounts of early court activities in the books. In addition, the *Nihon shoki* contains the texts of some

Japanese written records

200 poems, many of which seemed to have been derived from the oral musical tradition.

Predominant musical traits. It is apparent that by the 8th century the documentary history of Japanese music had begun. Although this claim predates an equal state of Western music history by some 100 years, certain interesting parallels between the two traditions can be made. Both seem more clearly established in the same general 200-year period, a short time when compared with Chinese music studies outlined above. Both developed a musical nomenclature heavily influenced by the music of religious organizations: the Roman Catholic church in the West, Buddhism in Japan. Both traditions were equally influenced by theories of a foreign culture from over the nearest sea: Greece in Italy and China in Japan. Herein many differences arise, one of the most significant being that, in the Japanese case, the foreign tradition of China at the time of its first major influence was alive and strong and could apply practical musical information and instrumentations as well as theories, whereas the Greek tradition was long dead by the same period, when the European monks turned to it for guidance. Nevertheless, one can see that the general length and beginning of each history is comparable. Before discussing Japanese music in chronological detail, an attempt should be made to envision general characteristics, realizing that in doing so the tendency is to apply aphorisms to music that stretches over a series of styles as old and varied as the music of Europe from Gregorian chant through Claude Debussy. With that caveat, general guidelines for the appreciation of Japanese traditional music can be put forth.

Aesthetic and formal ideals. These guidelines fall under three general concepts: (1) the sound ideal, (2) the structural ideal, and (3) the artistic ideal; but these three things are not clearly separate in any musical event.

In general one can say that the most common sound ideal of Japanese music is to produce the maximum effect with a minimum amount of material. For example, the *taiko* drum of the nō drama consists of a barrel-shaped body over which are lashed two cowhide heads some 20 inches in diameter stretched over iron rings. Wooden sticks are used to hit one head. Obviously, the sound potentials of the drum are many, but they are deliberately suppressed. For example, the sticks are made of very soft wood, and the strokes are applied only to a small circle of soft deerskin in the centre of the head. The *taiko,* like Japanese ink paintings, accomplishes a great deal by concentrating on very carefully chosen limitations of the medium.

Instrumental potential and tone colour

Another feature of much Japanese traditional music could be called the chamber music sound ideal. No matter how large an ensemble may be, one finds that the various instruments are set in such a way that the timbre, or tone colour, of each can be heard. This can be understood in Western chamber music and contrasts with the Western orchestral sound ideal, in which the primary intention is to merge all the instrumental sounds into one glorious colour. The colour separation of Japanese music is quite evident in the large court ensemble (gagaku), as well as in drama music and actual chamber ensembles such as the *sankyoku,* for koto (zither), samisen (plucked lute), and the end-blown *shakuhachi* flute. Such textures support the strong multilinear (as opposed to harmonic) orientation of East Asian music mentioned earlier.

The Japanese tripartite form

The structural intents of Japanese music are as varied as those of the West, but one of special interest is the frequent application of a three-part division of a melody, a section of a piece, or an entire composition. This is in contrast with the more typical two-part division of Western music. Of course, examples of both ideals can be found in the music of both cultures; the concern here is with broad generalities. The fundamental terminology of the Japanese tripartite form is *jo-ha-kyū,* the introduction, the scatterings, and the rushing toward the end. A Western musician might wish to compare this with sonata allegro form and its three parts (exposition, development, recapitulation). But the Western example relates to a complete event and involves the development of certain motives or melodic units (such as first and second theme), whereas the Japanese concept may be applied to various segments or complete pieces that are generally through-composed (*i.e.,* with new material for each segment). Japanese music reveals its logic and its forward motion not by themes but by a movement from one section to another different one until the final section is reached. Forward motion in motive Western music was often derived during the classical periods from the tension created by chord progressions. In Japanese music, such sonic events generally are not used. Nevertheless, the need for aurally recognizable patterns falling into a progression that the informed listener can anticipate is necessary in all music. In Japan such stereotyped patterns are melodic or rhythmic, not harmonic. They will be discussed in detail later; but the recognition, whether intellectual or aural, of the existence of such recurring patterns is essential to the appreciation of any music.

Word orientation. One of the artistic ideals of Japanese music is equally clear in all of East Asia. It is the tendency for much of the music to be word-oriented, either through actual sung text or through pictorial titles to instrumental pieces. With the exception of variation pieces (*danmono*) for the Japanese koto, one can seldom find a purely instrumental piece in the spirit of, for example, the Western sonata or symphony. Japanese ensemble pieces, like those mentioned earlier in China and Korea, are either dance pieces, instrumental versions of songs, or descriptive. This ideal in all of East Asia was not weakened until the late 19th century, when such music was forced to compete with Western idioms.

Guilds. By the same token, the ideal of the composer as genius, so dear to 19th- and 20th-century Western hearts, had little place in earlier East Asian music. In Japan, as in China and Korea, the names of many composers are known, but the actual setting of their music was and still is often done by a group of fairly anonymous people. One may know who was helping out at a given time and in a given place; but in any written form of the music their names, or even the name of "the" composer, may often be missing. The process might best be called communal composition. In the Orient, particularly in Japan, the performer is often the person remembered and noted. Such an ideal is understood in the West by fans of popular music. Although this ideal has given way to the Western composer "star" system in modern Japan, it does depict an important social setting for any appreciation of the older Japanese classical traditions. In keeping with this artistic ideal, one should add that often there is not one "correct" version of a given piece. Most traditional music is organized under guild systems, and thus each guild may have its "secret" version of a well-known piece. A given guild will play its version precisely the same way in each performance, for improvisation has practically no role in any of the major genres of all East Asian music. Differences are maintained between guild versions, however, in order to identify a given group's musical repertoire as separate from all the rest.

Manner of handling instruments

The separation of guild styles can be carried further to one more artistic ideal, which holds that it is not just what one plays on an instrument, it is how one plays it. For example, in the case of the *taiko* drum mentioned above, the manner in which a player sits, picks up the sticks, strikes the drum, and puts the sticks away will reveal the name of the guild to which he belongs and also can be used to judge his skill in performance. No Japanese instrument is merely played. One could almost say that its performance practice is choreographed. Such distinctions exist in the music of other East Asian cultures as well, although the clues to their understanding have not yet been revealed to outside listeners and viewers. This brief discussion of their existence in Japanese music will serve to enhance the appreciation of at least one Asian tradition as the discussion turns to a chronological study of its many styles.

The Nara period. *Codification of court music.* The previously mentioned documents from the Nara period (710–784) demonstrate how very active music was in the newly established capital in Nara. The general term for court orchestra music, gagaku, is merely a Japanese pronunciation for the same characters used in China for *ya-yüeh* and in Korea for *a-ak.* As Japan absorbed more and

more of the outside world, the music of the court, like that of T'ang dynasty China during the same general centuries, received an increasing variety of styles. In 702 these styles were organized under a music bureau (*gagakuryō*), and by the early 9th century an additional Outadokoro (Imperial Poetry Bureau) was created for handling Japanese-composed additions to the repertoire. Among foreign genres, the musical styles of the nearby Three Kingdoms of Korea have already been shown to be some of the first imports, Silla music being called in Japanese *shiragigaku,* Paekche music, *kudaragaku,* and Koguryō music, *kōkurigaku.* Music from the Three Kingdoms was sometimes called collectively *sankangaku.* Under all these terms were found still other Chinese and northern Asian traditions, in addition to music purported to have come from India as early as 736. Evidence of such a distant import can be found in a surviving court dance (*bugaku*) called "Genjō-raku," whose story about the exorcising of a snake can be traced to an ancient Indian Vedic tale. The date of 736 is also assumed for the entrance of music from Indochina, which survived for several centuries in a form of music called *rinyūgaku.* Although this tradition is now lost, there are extant detailed pictures of the ensemble along with other ancient instruments and a variety of dances in sources such as the 14th-century copy of the 12th-century *Shinzeigakuzu* scroll.

Influence of T'ang dynasty China. The dominant musical style of early gagaku was, naturally, from China and was called T'ang music (*tōgaku*). In Japan, as in Korea, the establishment and maintenance of such a music has made it possible for modern listeners to hear foreign versions of famous pieces long forgotten in the country of their origin. For example, there are names of pieces played and dances performed in Japan that are also found in T'ang Chinese lists. Unlike in China, however, many of these works are still played in Japan, and a few of the original costumes and masks used at that time are preserved. Perhaps the most valuable treasure in Japan for such materials from the ancient traditions of all of East Asia is the Shōsō-in, a storehouse built for the household goods of the emperor Shōmu after his death in 756. In this collection (which includes a few later additions from temples) one can find some 21 percussion instruments, 12 strings, and 12 winds, in addition to dance masks, notation, and drawings. Some of the materials are Chinese or Korean imports, while others are Japanese-made. The Chinese variant of the arched harp of the ancient Middle East (in Japanese the *kugo*) is best preserved here. The very decorations of certain instruments can also be historical gold mines. For example, the protective cover across the face of one plucked lute (*biwa*) contains the picture of a performer riding a camel near a palm-treed oasis. Another such cover depicts a group of foreign (*i.e.,* not East Asian) musicians accompanying an energetic dancer, all on the back of an elephant. Etchings along a hunting bow show scenes of dancing and music performances connected with a popular imported art of acrobatics and juggling called *sangaku.*

Instruments preserved in the Shōsō-in

THE HEIAN PERIOD

Music of the left and of the right. Further images of Japanese musical life can be captured from the Heian period (794–1185). In the very first chapter of the 10th-century *Ochikubo monogatari,* one of Japan's earliest novels, the sad fate of the heroine is noted by the fact that she was never able to learn how to play the Chinese seven-stringed *ch'in* zither, although she did have some training in Japanese koto music. The famous 11th-century works, such as Murasaki Shikibu's *Genji Monogatari* (*The Tale of Genji*), are filled with romantic koto, *biwa,* and flutes, as well as gagaku and *bugaku* performances and the singing of many songs. Diaries also show that the courtiers, now moved to Kyōto, found music to be a useful and frequent adjunct to their insular courtly life. It was in this period that the many forms of official court music were organized into two basic categories. The so-called music of the left was called *tōgaku* and contained the Chinese- and Indian-derived pieces. The music of the right was called *komagaku* and contained all Korean and Manchurian examples. In both categories there were pieces that by this time may have been Japanese arrangements or original compositions. The terms left and right were derived from the Confucian-based administration system of the new capital, which divided the entire government into such categories. In *bugaku* they controlled the costumes of the dancers, left dances emphasizing red, right dances, green. In gagaku these two major divisions standardized the instrumentation of the ensembles. When playing dance accompaniments, stringed instruments were deleted, but the two orchestras for purely instrumental performances were complete. Each used plucked 12-stringed zithers with movable bridges called *gaku-sō* or by the generic term koto. The string section was completed by a four-stringed plucked lute, the *gaku biwa.* A small hanging gong (*shōko*) and a large hanging drum (*tsuri daiko*) were found in both. The leader of a *tōgaku* piece would use a barrel drum (*kakko*) with two lashed heads struck with sticks, while a *komagaku* piece would be led by an hourglass *san no tsuzumi* drum similar to the Korean *changgo.* The standard melodic instrument for both was the double-reed *hichiriki,* with a *komabue* flute being added in *komagaku* and a *ryuteki* flute in *tōgaku.* The Japanese *shō* mouth organ appears in both.

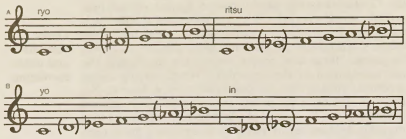
Dance costume and instrumentation

Musical notation. In modern performances the *shō* plays a fascinating cluster of harmonies, although there is some feeling that the original Chinese interpretation of its notation was melodic, with little or no harmonic addition. Part of the performance problem, outside the impressive age of the music, is that gagaku notations came only in part books, which were often rather like Western "lead sheets"; *i.e.,* they served as memory aids rather than detailed guides. For example, the *shō* notation gives only one note for every four beats of a standard piece. The modern interpretation is that each note represents the bottom of a chord, but the notes might actually have been the skeleton of a melody. The earliest surviving form of instrumental notation in Japan is a book, dated 768, of lute (*biwa*) music found in the Shōsō-in. The earliest flute part book is dated 966, and a few additional wind or string books remain from the 10th to the 13th centuries. More frequent sources can be found from the last 300 years. Scores did not exist until modern times.

The traditional part-book notations reflect the importance of oral, rote learning and the guidance of a teacher. For example, flute and *hichiriki* notations in their standard forms consist of columns first marked off by dots representing major percussion time markers (usually every four beats, though there are five- and six-beat pieces). Next, one finds a column of syllables called *shoga,* which were used to help one memorize the instrumental part by singing it. With this system it was even possible to substitute a vocal rendition of one part in an ensemble if that instrument was missing. Finally, there is another parallel column that contains a skeleton of notes or fingerings on the instrument itself. Obviously one can comprehend fully the subtle ornamentations and nuances of any melody notated in this manner only through the guidance of a teacher. In the case of the string notation, one generally finds only the names of stereotyped patterns along with occasional notes. The percussion notations likewise consist of names for stereotyped patterns. If both the strings and percussion are played as written, they appear to be merely time markers, in accordance with the colotomic principle found in much Southeast Asian and some East Asian music (the demarcation of time intervals by the entrance of specific instruments in prescribed order, a procedure contributing to the musical structure and imparting a sense of progression).

But, as in jazz, there must have been more to the music hidden in the oral tradition. Further clues as to the performance practice of this music in addition to its underlying music theory and its practical uses are found in several important sources. In 735 an ambassador, Kibi Makibi, brought back from China a 10-volume digest of musical matters (called in Japanese *Gakusho yoroku*), which implies the Chinese foundation of the art. In 1233 a court dancer, Koma Chikazane, produced another 10 volumes—the *Kyōkunshō,* describing Japanese gagaku matters. Of equal value is the *Taigenshō,* written by a

IX



gagaku musician, Toyohara Sumiaki, in 1512, when court music seemed on the verge of extinction.

Tonal system. By a combination of these sources—Buddhist music-theory tomes, part books, and present-day performance practice—it is possible to understand many of the basic principles upon which ancient Japanese music was founded. From what has already been said about the beginnings of Japanese court and religious music, it is not surprising to find that the complete tone system of both consists of the Chinese 12 tones shown in notation I, the only difference being the Japanese pronunciations of the characters for each pitch name. The scales in IX-A show that Japanese ancient music followed the East Asian tradition as well in the use of two seven-tone scales, each with a pentatonic core. The *ryo* scale (set on C for the sake of comparison) shows no great difference from the Chinese scale in III; but the *ritsu* scale seems to reveal the early presence of an indigenous Japanese tonal ideal with the placement of its half steps.

Principal tonalities

Japanese gagaku and Buddhist music theories contain most of the classical Chinese ideas concerning transpositions and modes, but in practice the two scales shown in IX-A could be constructed on only three pitches each: *ryo* on D (*ichikotsu*), G (*sōjō*), and E (*taishiki*); and *ritsu* on E (*hyōjō*), A (*ōshiki*), and B (*banshiki*). Note that the pitches for such transpositions form a classic pentatonic (D–E–G–A–B). The two names for the pitch E are present in order to make a distinction between the two scales possible on that same tone. In the unaccompanied court songs and the chants of Buddhism, one can observe the use of other transpositions, for all oral traditions in the world "adjust" notated pitches to the preferences of given singers. In gagaku instrumental music, the six tonalities are observed, part books for each instrument being organized in sections by the tonalities of the compositions. A few pieces are found in more than one tonality. A transcription of part of the basic melody for such a composition, "Etenraku," is shown in notation X-A/B. Although set in two *ritsu* tonalities (*hyōjō* and *banshiki;* X-C), it is obvious from this example that the piece, which is a "crossover" (*watashimono*), is more than merely transposed.

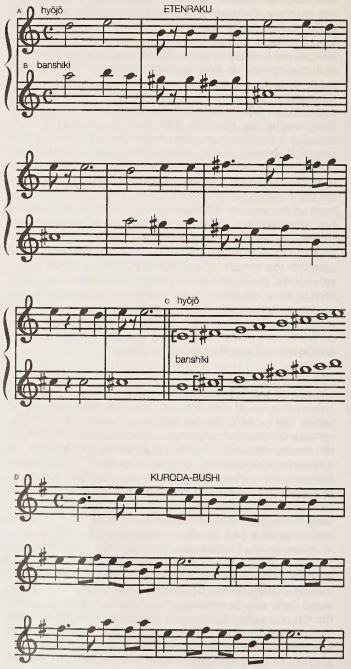
Although the scale has been transposed, yet the pitch centre of the melody also has been changed (from E to C♯). This is one of the few clear examples in performance practice of the mode systems spoken of in music theory. A glance at the related folksong "Kuroda-bushi" (X-D; see below) and at the *in* scale (IX-B) provides a preview of an emphasis on such a different mode centuries later.

Vocal music. In the poetry-oriented court life of Japan, secular vocal music would obviously be important. Many of the poems in classical collections seemed originally to have been song texts. One of the oldest secular song forms is *saibara,* which was first inspired by the singing of pack-train drivers. Among the new fads of Heian period vocal music (called collectively *eikyoku*) were *rōei,* songs based on Chinese poems or imitations of them, and *imayō,* contemporary songs in Japanese. Many gagaku melodies were given texts to become *imayō* songs, while others were derived from the style of hymns used by Buddhist missionaries. Little of these vocal traditions remains, but memories of their importance are preserved in nearly every novel and diary of the period. For larger surviving repertoires it is necessary to turn to religious music.

Shintō music. The indigenous religion of Japan, Shintō, was closely connected with the legendary legitimacy of the emperor. Thus, special Shintō music was devised for use in Imperial shrines, a tradition already familiar from the discussion of China and Korea. In Japan such Shintō music is called *kagura.* The kind of music and ritual used exclusively in the Imperial palace grounds is called *mi-kagura,* that in large Shintō shrines, *o-kagura,* and Shintō music for local shrines, *sato-kagura.* The *suzu* bell tree, mentioned before as among the earliest known Japanese instruments, is found in all such events; and the equally ancient *wagon* zither can be heard in the palace rituals and sometimes in the larger shrines.

General Shintō chanting (*norito*) is rather straightforward, whereas the surviving music of *mi-kagura* is more complex. Unison choruses of men are accompanied by the *hichiriki* oboe, a *kagura-bue* flute, the *wagon* zither, and the periodic rhythmic markings of a pair of long, thin *shaku byōshi* clappers. The music for *mi-kagura* ceremonies is divided into two types: one to praise the spirits or seek their aid (*torimono*), the other to entertain the gods (*saibari*) in the tradition, mentioned earlier, of the mythological amusements given before the sun goddess. Perhaps the most famous surviving dance suite from the Shintō tradition is *Azuma asobi* (*The Entertainment of Eastern Japan*), which can be seen as a courtly reflection of the agricultural base of Japan in its annual performances during the spring equinox and the summer solstice. The work is said to be an imitation of the dance of a heavenly maiden who performed on the beach of Suruga in the 6th century. *Azuma asobi,* along with *bugaku* dances, may be seen at many other Imperial, national, and shrine occa-

X



sions—dim but nevertheless impressive reflections of the colourful courtly life of Japan of centuries ago.

Mi-kagura is exclusively a male event, but Shintō female dancers (*miko*) are found in other shrines. Historical documents show that the Heian court, like courts in ancient China or, for that matter, all over the world, appreciated the value of female dancers and their music. In later times the Heian-originated *shirabyōshi* female dancer-musicians became important elements in the transfer of courtly and religious traditions into later theatrical forms. The major source of religious musical influence is found elsewhere in the Buddhist temples.

Buddhist music. There are many forms of Buddhist hymns, such as *saimon,* as well as semireligious dance songs, such as *goeika, nembutsu odori,* and the *bon odori* performed to folk festivals. But the basis of Buddhist classical music and hence the core of Buddhist influence on Japanese art music is found in the theory and practice of chanting known generically as *shōmyō.* Such a tradition came originally from foreign Buddhist missionaries and next from Japanese converts studying in China. Noted sources from the many Japanese interpretations of this tradition are the *Shōmyō yojinshu* by Tanchi (1163–1237) of the Tendai sect and the *Gyosan taikaishu* (1496) of the Shingon sect. The theoretical bases of these studies are similar to the ones already discussed under the topic of gagaku. Here need be added only comments about Buddhist notation systems. Most early chant notations used neumes, squigglelike signs that, like those of the early Christian traditions, served primarily as memory aids with which an initiate could recall the details of a given melody. The most influential system was the so-called *go-in hakase,* attributed to Kakui (b. 1236) of the Shingon sect. Under this method the five notes of each of three octaves of a pentatonic scale were indicated by the angle of a short line, rather like the hands on a clock. Variations of this method were of great influence in the notation of all vocal music of the period and continue to be used in Buddhist chant today.

Systems of notation

KAMAKURA, MUROMACHI, AND TOKUGAWA PERIODS

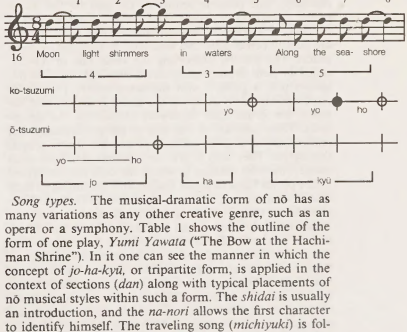
Nō music. The Kamakura period (1192–1333) marks the end of Heian court splendour and the start of a new military government located in Kamakura, far away from Kyōto. In such a context it is not surprising to find the development of long narratives of military history and the flourishing of plebian theatricals. The story of the defeat of the Heike clan (the *Heike monogatari*) was known in mansions, war camps, and temple grounds primarily as sung by *biwa*-playing bards. As in the traditions of ancient Greece and Europe, these minstrels were often blind or built their style in that of the blind-priest lute tradition (*moso biwa*) in which mendicant monks used to recite sutras (scriptures) from house to house or at temples. More lucrative forms of entertainment grew under the circus acts that developed out of the *sangaku* (folk theatricals) mentioned above; its companion comic acts, *sarugaku* (literally, monkey or mimic music); and theatricals derived from folk rice-planting dances, *dengaku.* Street parades (*fūryū*) and Buddhist entertainments (*ennen*) also were part of the colourful scene. By the subsequent Muromachi period (1338–1573) the terms *sarugaku-no-Nō* and *dengaku-no-Nō* had become the dominant terms for temple and shrine pantomime and dialogue dramas, while the comic interludes of such plays were called *kyōgen.* Through the support of the military rulers and the efforts of individual artists such as Kan'ami (1333–84) and his son, Zeami (1363–1443), the first major form of Japanese theatre developed. It became known eventually as nō (see below *Dance and theatre: The development of dance and theatre in the East Asian nations*).

The music of nō as it is performed today consists of vocal music (*yōkyoku*) with an instrumental ensemble known collectively as the *hayashi.* The singing is done by the actors or by a unison chorus (*jiutai*). The four instruments of the *hayashi* are a flute (*Nō-kan*), the *taiko* stick drum described earlier, a small hourglass drum (*ko-tsuzumi*) held on the right shoulder, and a larger one (*ō-tsuzumi*) placed at the left hip.

Melodic principles. The writings of Zeami, such as the *Kaden-sho,* contain terms reflecting the traditional tone systems and terminologies of former times. A distinction was made between the recitative section (*kotoba* or *serifu*) of a play and melodic parts (*fushi*). The melodies of nō can be categorized into two basic styles, the strong (*tsuyogin*) and the lyric (*yowagin*). Their differences are most evident in the placement of fundamental tones and the use of auxiliary tones around them. In the lyric style the three basic tones (*jō, chū,* and *ge*) are a fourth apart (see notation XI-C). The movement to and from each note is regulated in a manner comparable to the regulated approach to certain intervals in 16th-century Western counterpoint. Similar but different laws are applied to the strong style, the fundamental tones of which are much closer, the standard procedure today being that *jō* and *chū* are one pitch and *ge* is approximately a minor third below (XI-B). Melodies contain various formulas and ornamentations the names of which often reflect earlier traditions in Buddhist chant and court secular songs. The choice of combinations depends on the musical needs of the given dramatic text as well as the position of the music in the general form of the piece. Once more, such restrictions in an "exotic" music should seem reasonable to students of Western art music with its needs for stylistic restraints. The notation of nō singing (sometimes called *utai*) is derived from simpler Buddhist and early *biwa* forms that used teardrop-shaped neumes along with important pitch names to remind singers of the performance practice of a given passage. This so-called sesame-seed notation (*gomaten*) remains basic to nō vocal music today, and there are many detailed books in modern Japanese to help the initiate follow the music with the aid of a teacher. Variations in notation style and in the interpretation of specific passages are maintained by the various schools of nō along with the "secret piece" tradition basic to much of Japanese traditional music since its beginnings.

Fundamental and auxiliary tones

XI



Song types. The musical-dramatic form of nō has as many variations as any other creative genre, such as an opera or a symphony. Table 1 shows the outline of the form of one play, *Yumi Yawata* ("The Bow at the Hachiman Shrine"). In it one can see the manner in which the concept of *jo-ha-kyū,* or tripartite form, is applied in the context of sections (*dan*) along with typical placements of nō musical styles within such a form. The *shidai* is usually an introduction, and the *na-nori* allows the first character to identify himself. The traveling song (*michiyuki*) is followed by a song emphasizing higher tones (*ageuta*). Songs

Table 1: Tripartite Form Concept (*Jo-ha-kyū*) in the Nō Drama *Yumi Yawata*

sections (*dan*)	placement of musical styles
Jo (*mae-dan*)	*shidai na-nori michiyuki* (*ageuta*)
Ha	
First *dan*	*issei sashi sageuta ageuta*
Second *dan*	*kotoba sashi ageuta kotoba*
Third *dan*	*kuri sashi kuse rongi*
Kyū (*ato-dan*)	*ageuta deha sashi issei mai rongi kiri*

with lower melodic emphasis (*sageuta*) and other styles (*issei, kuri,* and *rongi*) mix with more recitative sections (*sashi, kotoba*) and with entrances (*deha*), closings (*kiri*), and dances (*kuse* and *mai*).

Within each of these sections and subsections one must remember that drama and text have their influence as well. Although the drama is not all poetic, the earlier discussion of Chinese theatre should prepare one for the fact that many lines of the text are indeed actually poems or are influenced by poetic form. Although there is great variety in the syllable lengths and combinations, the most common divisions in nō are into lines of 7 or 5 syllables. Such a 12-syllable line has been constructed in English in notation XI along with a 16-syllable variant so that the reader can compare certain basic principles in the settings of Japanese nō texts. The placement of a text rhythmically can be done in three major ways: *o-nori,* with 1 syllable per beat (XI-A); *chū-nori,* with 2 syllables per beat (XI-B); and *hira-nori,* in which 12 syllables are worked into an eight-beat frame (XI-C). The setting shown in XI-A is in the spoken (*kotoba*) style, XI-B, in the strong (*tsuyogin*) singing, and XI-C is lyric (*yowagin*).

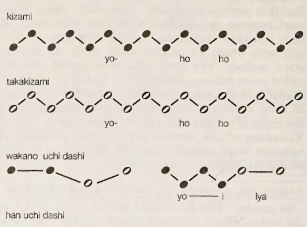
Setting of texts

Note in XI-C that, although there are two sections of 7 and 5 syllables, the actual division of their presentation (4 + 3 + 5) is tripartite—*i.e.,* in the form of *jo, ha,* and *kyū.* Lines of greater numbers of syllables in *hira-nori* require more beats, and 12-syllable lines themselves can be handled in many other ways. Thus, one must not take this artificial example to be representative of the only style of nō text setting any more than one can consider a given eight-beat phrase from one Mozart opera as being sufficient to show all there is to aria form. Nevertheless, the settings given in notation XI show in their syllable displacements and melodic contours that such "Oriental" music, like that of all cultures, does have a thorough internal logic. Nō music shares with Western art music the extra convenience of a complete written music theory, of which this short example demonstrates only a few elementary principles. The principles may be taken one step further by constructing a typical drum accompaniment for the *hira-nori* version of the texts.

Function of drum patterns. The lower portion of XI-C represents one possible setting of the text by the music of the two *tsuzumi* drums. The music of the nō drums consists of a series of named, stereotyped patterns that are aurally perceivable and that tend to progress in given orders. The pattern supporting the vocal line in XI-C is called in both drum parts *mitsuji,* although they do not always play a pattern with the same name. The circles represent moments in which the drum is struck, and the words are drum calls (*kakegoe*) uttered by the drummers. These calls are as essential to the performance and recognition of a given pattern as are the drum sounds themselves. They help to give each pattern a unique aural image. Also, the manner in which the calls are performed by the player helps to signal and control the timing of each beat in a music that is often very elastic in rhythm. Note in addition that, in XI-C, the first call of each drummer builds up to his first actual striking of the drum, which in turn marks the divisions between the three parts—*i.e.,* the *jo, ha,* and *kyū* phrasing of the vocal line. In actual performance the spacing of the beats in the vocal line may not be even with those of the drums, though beats 3 and 5 often match. The controlled but subtly varied relations of song to rhythmic accompaniment in nō drama are analogous to harmonization in Western music.

This point may be taken one step further by showing (in notation XII) a set of named patterns for the *taiko* stick drum, an instrument used only in dance sections of a play. Placement of the dots shows right- and left-hand strokes; black dots indicate the softer—and light dots the louder—strokes. The few patterns (*tetsuke*) in the example are from a set of 59 found in a *taiko* instruction book. A study of them shows that they belong to families (*tegumi;* the *kizami* family in A and B and the *uchi dashi* group in C and D). Once more the principles of harmony in Western music come to mind; *i.e.,* a C-major and C-minor chord are related because they share some common aural traits (the tones C and G in this case). If one looks into a lesson book of *taiko* music, one will find that, as in many Western harmony books, the student will be told what patterns may appear before or after each item. In Western music one learns in a similar way that certain chords may be preceded and followed by others. In both the Japanese and Western cases there is a selection of permissible choices of earlier or following events for each pattern. Thus, it would seem that the concept of prediction and anticipation is fundamental to the listener's sense of logical progression in the music of both. The major difference in this case is that the aurally perceivable, named, stereotyped pattern of Western traditional music is a vertical sonority called a chord, whereas in nō music it is a horizontal time unit called a *tetsuke.*

XII



Role of the flute. The music of the nō flute (*nō-kan*) has many stereotyped patterns, but it functions in rather different ways from the drums of the *hayashi.* Although it seems originally to have related to the vocal line, today it does not play the singer's melodies. The nō flute looks like the *ryuteki* flute of gagaku; but a short cylinder has been inserted inside its tube so that the upper holes of the flute overblow a seventh (as A–G) rather than the octave, as with all other flutes. This unique characteristic of the flute seems to reflect the tonal principles of nō mentioned earlier, whose basic notes outline a seventh (A–D–G in notation XI-C). The flute may give a specific pitch to set the tonality of a vocal entrance, but its normal functions today are to signal sections of the form and to play one of the dozen standard dance pieces used in various nō dramas. The dance pieces are set in sections (*dan*). Each piece is learned by rote or by a notation consisting of flute mnemonics placed in a frame of eight squares representing beats in a manner rather like the Korean notation mentioned earlier. The flute's seven holes are fingered by the middle joint of the fingers instead of the tips, producing an impressively fluid melody that would not fit into the graphic notation system of traditional Western music. At the same time, it does not compete with the nō vocal line by being too melodically clear. Indeed, a discussion of the nō flute seems to be a particularly appropriate finale in this brief survey of nō music (*nō-gaku*); for it shows that the form of each musical item—whether performance practice, pitch relations, structure, or notation—follows the needs of its functions. The music of nō drama seems on first hearing to be one of the most puzzling of East Asia's exotic sounds, but a study of its principles can make it become as reasonable and as beautiful as a Bach cantata.

Koto music. *Schools and genres.* The 13-stringed zither with movable bridges called the koto has been mentioned as one of the basic instruments of the court ensembles as well as a common cultural accoutrement for court ladies. The development of independent solo and chamber music genres for this instrument becomes more evident as one moves into the Muromachi period. The earliest surviving school of solo koto music is Tsukushi-goto. It was first noted on the island of Kyushu in the late 16th century where, over the centuries, court refugees and exiles gathered during upheavals in Kyōto. Earlier Chinese influences also are claimed as part of its creation, though historical facts are obscure. Tsukushi-goto repertoire is said to begin with variants of *imayō* court songs. Sets of songs were accompanied by the koto and sometimes by the three-stringed plucked samisen (*shamisen* in Tokyo dialect). The sets were called *kumiuta,* a term applied to much of the chamber music that followed. The 16th-century priest Kenjun is credited with the creation of the school and its first compositions. The tradition became more secular when it appeared in Edo. There, a 17th-century blind musician named Johide, claimed as a student of Hosui, a student of Kenjun, developed his own version of such music. He added compositions in more popular idioms and scales, named himself Yatsuhashi Kengyō, and founded the Yatsuhashi school of koto. The title Yatsuhashi was adopted later by another apparently unrelated school to the far south in the Ryukyu Islands.

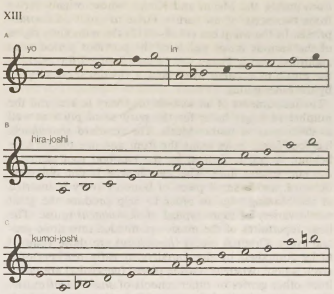
Additional schools of popular, or "vulgar," koto (*zokuso*) reflected the mercantile life of the new Tokugawa (also called Edo) period (1603–1867). In 1695 another third-generation extension of the Kenjun's koto tradition was Ikuta Kengyō, who began his Ikuta school. The term *kengyō* had been one of the basic ranks of musicians under the guild system and so is frequently found in professional names, but the name Ikuta remained as one of the primary sources of koto music until the creation of still another school by Yamada Kengyō (1757–1817). In present-day Japan the Ikuta and Yamada schools remain popular, whereas the earlier traditions have faded considerably. Both schools have provided famous composers; and there are several pieces from their schools, as well as a few earlier works, that are now shared by the guilds as part of the classical repertoire of the koto. The slightly longer and narrower shape of the Ikuta koto produces a tone easily distinguishable from that of the Yamada school.

Koto music is known in general as *sōkyoku.* In the koto solo instrumental music (*shirabemono*), the most important type is the *danmono,* a variation piece in several sections (*dan*), each normally of 104-beat length. The term for koto chamber music, *sankyoku,* means music for three. The standard instrumentation today consists of a koto player who also sings, along with performers on a three-stringed plucked samisen lute and an end-blown *shakuhachi* flute. In earlier times a bowed variant of the samisen called the *kokyū* was used more often than the flute. The basic genre of chamber music is called *jiuta* and combines the earlier *kumiuta* tradition of accompanied song with instrumental music by alternating sections with singing (*uta*) and instrumental interludes (*tegoto*). After the 19th century a second embellishing koto part (*danawase*) often was added to the instrumental interludes. Innovations carried out during the 20th century will be covered later in this article.

Tunings and notation. Each school of koto music from the courtly tradition to the present time involves changes in the structure of the instruments as well as changes in playing method and notation. The ancient court koto (*gaku-so*) is similar to the modern koto and is played with picks (*tsume*) on the thumb and first two fingers of the right hand or with bare fingers, although, unlike the Ikuta and Yamada styles, the left hand is not used to alter the tone by pressing the string on the other side of the movable bridges. Its notation consists primarily of the names of basic patterns in addition to occasional melodic fragments and the text. The survival of such music is dependent on a continuing viable rote tradition; thus, most of the tradition is lost.

The tunings of the 13 strings of the court koto were derived from the modes of the *ryo* and *ritsu* scales of the earlier periods. The tunings used in the Edo koto traditions, however, reveal new, apparently indigenous, tonal systems. These concepts were eventually categorized under the two scales called *yo* and *in,* shown in notation XIII-A. The tunings in XIII-B and XIII-*C* reflect the new kinds of pentatonism of the period with their use of half steps.

XIII



The *hira-joshi* tuning appears in such famous early works as *Rokudan* (*Six Dans*) ascribed to Yatsuhashi Kengyō, the "founder" of the modern koto styles. In all, there are some 13 standard tunings for the koto and many variants. Like all the other popular music in Japan from the 17th century on, these koto tunings are based either on the older tradition preserved in part in the *yo* form or on the more "modern" *in* scale. One can note in the 19th century occasional pieces deliberately written in the previous gagaku mode style as well as the use of the Holland tuning (*oranda-choshi*), the Western major scale derived from the Dutch business area on Deshima in Nagasaki. Nevertheless, the *yo-in* system remains the fundamental tonal source for new Japanese music from the 17th century on, exceptions being revived court music, new nō plays, and the work of avant-garde composers after World War II.

Printed collections

The earliest printed notations of koto, samisen, and flute pieces from the Tokugawa period are found in the *Shichiku shōshinshū* (1664), the *Shichiku taizen* (1685), and the *Matsu no ha* (1703). Although many sections of such collections contain only the texts of songs, one can find certain pieces that parallel the line of words with numbers representing strings on the koto or finger positions on the samisen, names of stereotyped koto patterns, or mnemonics for the particular instrument with which the piece is learned. In the late 18th century both the koto and the samisen traditions developed more visually accurate notations. The koto version (first seen in the *Sōkyoku taisho,* 1779) used various-size dots to indicate rhythm. In the early 19th century, string numbers were placed in columns of squares representing rhythm, as in the system mentioned earlier in Korea. The numbers and squares eventually were combined with the ♩ bar-line concept of the West; so that the notations of both schools today, although separate systems, maintain a balance of traditional and Western ideas. Their modern compositions attempt to do the same as well; but before they can be treated, attention must be given to the traditions connected with the other major instruments of the Tokugawa period.

Schools of shakuhachi flute music. The *shakuhachi* end-blown flute is a variant of the Chinese *hsiao,* and examples of it can be found in the famous 8th-century Shōsō-in treasure house mentioned earlier. During the Muromachi period (1338–1573) a smaller Japanese version called the *hitoyogiri* became popular as a solo instrument, but the best-known form of the *shakuhachi* is the one developed in the Tokugawa period. The instrument was used by *komusō,* priests who begged or sometimes spied while wandering through the streets playing the flute incognito, their heads covered by a special wicker

basket hat. With the changes in contemporary Japanese society, many former warriors no longer carried their swords, whereas young merchants carried more money. One curious side effect of such changes was the occasional appearance of a *shakuhachi* tucked in the back of one's belt for use as a musical device or as a club.

The major schools of *shakuhachi* music today come from guilds, the Meian and Kinko, whose origins derive from two sects of an earlier Fukeshu guild of *komusō* priests. In the Meiji era (1868–1912) the monopoly rights of the various music guilds of the previous period were abolished; and a Tozan school was founded for teaching the music to amateur musicians, a custom soon adopted by the other guilds.

The instruments of all schools may vary in size and the number of finger holes for the purpose of pitch as well as differences in timbre ideals. The standard *shakuhachi* has four finger holes along the front and one thumb hole behind. A bell is formed by the bamboo root stems at the end of the flute. The mouthpiece is cut obliquely outward, and a small piece of bone or ivory is inserted at the blowing edge in order to help produce the great subtle variety of tones typical of *shakuhachi* music. The basic repertoires of the music are divided into three general types. Original pieces (*honkyoku*) are those claimed to be composed by the founders or early teachers of a given school, whereas outside pieces (*gaikyoku*) are taken from other genres or other schools of *shakuhachi* music. New pieces (*shinkyoku*) continually appear and are kept in that category. *Shakuhachi* notation varies with each school; however, all are based on mnemonics with which the music is taught. Given the exceptional subtlety of tone changes and ornamentation in all traditional *shakuhachi* music, such a notation system seems quite logical. The beautiful introverted sounds of *shakuhachi* music seem closer to Buddhist chant than to other instrumental forms and are best learned by the ear and heart rather than by the eye and brain.

Samisen music. The three-stringed plucked lute of Japan is known as the *shamisen* in the Tokyo area or as the *samisen* in the Kansai district around Kyōto. It seems to have arrived in Japan as an import of the *sanshin*, or *jamisen*, from the Ryukyu Islands in the mid-16th century. The Ryukyu form of the instrument, with its oval body and snakeskin, is obviously derived in turn from the Chinese *san-hsien*. Such an origin is reinforced by collections of early Ryukyu music, which use a so-called *kukunshi* notation similar to the Chinese symbols shown in notation V. The Japanese samisen underwent considerable physical change, its body being rectangular and the skins coming from a cat or dog. Apparently under the influence of contemporary *biwa* lute traditions, the plectrum of the instrument was changed from the talonlike pick of the Ryukyus to a wooden or ivory *bachi* with a thin striking edge. In addition, the lowest string was kept off the small metal upper bridge near the pegbox so that it produced a buzzing sound (*sawari*) distinctly reminiscent of the tone of a *biwa*. The three basic tunings of the Japanese instrument are *hon-chōshi* (b–e′–b′; b represents the B below middle C, b′ the B above); *ni agari* (b–f♯′–b′); and *san sagari* (b–e′–a′). These tunings have remained standard to the present day, although there are occasional variants.

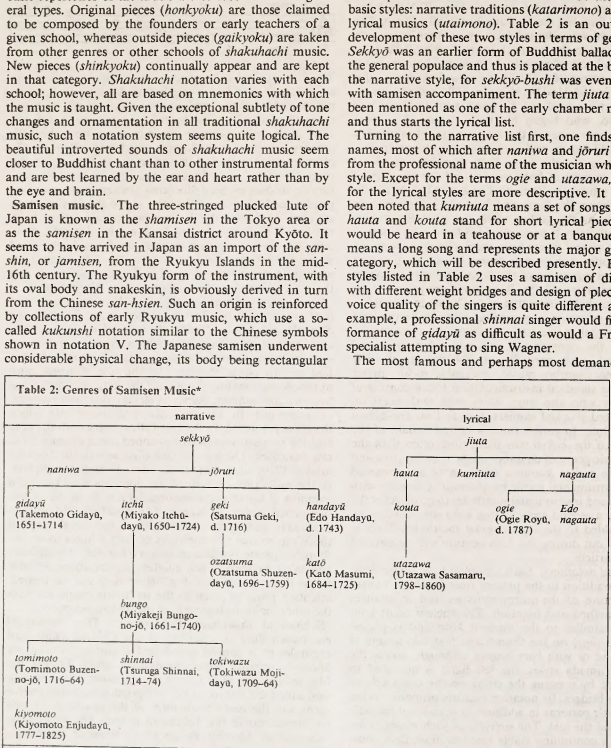
Greater variety is found in the many genres of samisen music. The earliest types seem to have been played by old *biwa* entertainers around Ōsaka, a city then called Naniwa; hence the name of the new genre was *naniwa-bushi*. Samisen was used for folk music and party songs, but, in keeping with the *biwa* origin of the first performers, narrative music was of prime importance. Such music became known as *jōruri*, the term being derived from the title of a famous story of the princess Lapis Lazuli (*Jōruri-hime monogatari*). As different guilds of samisen evolved, it was possible in modern times to divide them into two basic styles: narrative traditions (*katarimono*) and basically lyrical musics (*utaimono*). Table 2 is an outline of the development of these two styles in terms of genre names. *Sekkyō* was an earlier form of Buddhist ballad drama for the general populace and thus is placed at the beginning of the narrative style, for *sekkyō-bushi* was eventually done with samisen accompaniment. The term *jiuta* has already been mentioned as one of the early chamber music forms and thus starts the lyrical list.

Jōruri, or narrative, music

Turning to the narrative list first, one finds a mass of names, most of which after *naniwa* and *jōruri* are derived from the professional name of the musician who began the style. Except for the terms *ogie* and *utazawa*, the names for the lyrical styles are more descriptive. It has already been noted that *kumiuta* means a set of songs. The terms *hauta* and *kouta* stand for short lyrical pieces such as would be heard in a teahouse or at a banquet. *Nagauta* means a long song and represents the major genre in this category, which will be described presently. Each of the styles listed in Table 2 uses a samisen of different size with different weight bridges and design of plectrums. The voice quality of the singers is quite different as well. For example, a professional *shinnai* singer would find the performance of *gidayū* as difficult as would a French opera specialist attempting to sing Wagner.

The most famous and perhaps most demanding of the

Table 2: Genres of Samisen Music*



*Names of founders of some styles are given in parentheses.

narrative styles is *gidayū*, named after Takemoto Gidayū (1651–1714), who worked with Chikamatsu Monzaemon in the founding of the most popular puppet-theatre tradition (known as bunraku) of Ōsaka. The *gidayū* samisen and its plectrum are the largest of the samisen family, and the singer-narrator is required to speak all the roles of the play, as well as to sing all the meditations and commentaries on the action. The part is so melodramatic and vocally taxing that the performers are often changed halfway through a scene. There is little notated in the books (*maruhon*) of the tradition except the words and the names of certain appropriate stereotyped samisen responses. The samisen player must know the entire drama by heart in order to respond correctly to the interpretations of the text by the singer. The two musicians sit on a platform to the stage left of the theatre and through the intensity and skill of their performance help bring life and pathos into the wooden characters who move with frighteningly realistic gestures in the hands of three puppeteers. The power of *gidayū* is such that it can be heard in concert versions as well. In the 19th century a school of female performers (*onna-jōruri*) carried on the concert tradition with equal ability.

Music for the puppet theatre

Kabuki theatre. The *nagauta* form of lyric music, like most of the other narrative forms, began with a close relation to the kabuki popular theatre of the Tokugawa period. The first kabuki performances used instruments (*hayashi*) from the nō drama. Because kabuki was related to the flourishing demimonde of the major cities, however, the music of the party houses and brothels was soon added to the theatre. By the mid-17th century the names of *nagauta* singers and samisen players were listed on posters along with the cast. In the same manner, the names of musicians in many of the other genres listed in Table 2 were adopted to denote parts of a play. Although nearly all the music listed can be heard in concert forms today, the major genres still included in kabuki productions are *gidayū*, *tokiwazu*, and *kiyomoto* from the narrative styles, and *nagauta* from the lyrical. Rather than being discussed individually, they will be viewed in the total theatrical context and later brief reference will be made to their concert forms.

Onstage music. Kabuki as theatre is discussed below in the section *Dance and theatre: The development of dance and theatre in the East Asian nations.* Its musical events can be divided into onstage activities (*debayashi*) and offstage groups (*geza*). In plays derived from puppet dramas, the *gidayū* musicians, called here the *chobo*, are placed on their traditional platform offstage left or behind a curtained alcove above the stage-left exit. If other genres are used, the performers are placed about the stage according to the scenery needs of the play. There are some plays in which several different kinds of onstage music are required, a situation called *kake-ai*. The most common dance scene today, however, is one in which the onstage group consists of *nagauta* musicians and the nō *hayashi*. The samisen and singers are placed on a riser at the back of the stage, and the *hayashi* sit before them on floor level—thus, their other name of the *shitakata*, meaning "the ones below."

There are as many different types of dances that require different kinds of music as there are in Chinese or Western opera. In a general view, perhaps the most intriguing side of this variety is the relation of the older drum and flute parts to the vocal and samisen melodies of the Tokugawa period. In totally kabuki-style pieces, the *tsuzumi* drums play a style called *chirikara* after the mnemonics with which the part is learned. The patterns of this style follow closely the rhythm of the samisen part. If the nō flute is used as well, it is restricted to cadence signals; if a simple bamboo flute (*takebue* or *shinobue*) is substituted, it plays an ornamented (*ashirai*) version of the tune. There are many sections, however, in which the drum patterns and nō flute melodies discussed earlier are combined with samisen melodies. In a classical repertoire of hundreds of set pieces, there are many different combinations, but to many listeners these situations seem rather puzzling at first hearing, with apparently two kinds of music going on at the same time. If the situation is from a play derived from a former nō drama and uses the full *hayashi*, one notes first that the flute is not in the same tonality as the samisen nor is it playing the same tune. The drums in turn do not seem to relate rhythmically to the melody, as they do in the *chirikara* style. The drums and flute are, in fact, playing named stereotyped patterns normally of eight-beat length as in the nō. The essential difference between them and the samisen melody is that they do not seem aurally to have the same first beat. A given samisen melody will often make room through silence for an important vocal call in the drum patterns, but the deliberate lack of coordination of beat "one" creates a vital rhythmic tension that makes the music drive forward until it is resolved at a common cadence. Each part is internally rigid and progressive, but its conflict with the other parts forces the music (and the listener) to move the musical event through a time continuum toward a mutual completion.

Tension between simultaneous styles

The nō flute music is frequently related to the *taiko* stick-drum rhythm, so that they can be considered as a common unit rather than separate parts. There are situations in which the *tsuzumi* play *chirikara* patterns in support of the samisen melody, while the *taiko* and nō flute play either nō patterns or later kabuki-named drum patterns "out of synchronization" with the other music. At such moments one can see that in kabuki dance music, as in Western Classical music, there are three kinds of musical needs. In the West they are melody, rhythm, and harmony. In this music they are melody, rhythm, and a third unit of one drum and a flute that functions like harmony although its sound is totally different. If this third Japanese feature is called the dynamism unit, then it can be said that *nagauta* dynamism and Western traditional harmony both serve to colour the line, to create tension that drives the music onward, and to help standardize the formal design of the piece by clarifying cadences or by creating the need for them. All this brings back the earlier point that music is not an international language. The equally logical but different aspect of this music is certainly most obvious and striking.

The formal aspects of kabuki music are as varied as the plays with which music is connected. In dance pieces derived from nō plays, many of the sectional terms of the nō mentioned above are found. The classical kabuki dance form itself often consists of sections divided into the traditional tripartite arrangement as shown below:

deha or *jo*
oki michiyuki

chūha or *ha*
kudoki, monogatari, odori ji

iriha or *kyū*
chirashi, dangire.

Generally speaking, the *oki* represents all kinds of introductory instrumental sections (*aigata*, or in this case *maebiki*) or vocal parts (*maeuta*) before the entrance of the dancer. The *michiyuki* usually incorporates the percussion section as the dancer enters. The term *kudoki* is found in the early history of samisen music as a form of romantic music and is used here for the most lyrical section, in which the percussion is seldom heard. The *monogatari* (story) relates to the specific plot of the dance, and the *odori ji* is the main dance section, rather like the *kuse* or *mai* of the previous nō form. During this section, the bamboo flute may appear for contrast and, in nō style, the *taiko* drum may be important. The *chirashi* contains more active music, and the final cadence occurs during the *dangire*. There are endless variations and extensions of this form, but the many specific instrumentational and stylistic traits found in each of these sections help the listener to become aware of the logical and necessary progression of a given piece through a moment of time to its proper ending.

Most early collections (*shōhon*) of onstage music consisted of the text and samisen mnemonics (*kuchi-jamisen*, mouth samisen) of instrumental interludes (*ai-no-te*). In the 18th century some of the lyrical forms began to use syllables to represent fingering positions on the instrument, a system called the *iroha-fu*. In 1762 a set of circles

Changes in notation in the 18th century

with various extra markings along with the string number were combined in a book called the *Ongyoku chikaragusa* to create a more accurate if complicated system. Further rhythmic refinements were created in the 1828 *Genkyoku taishinsho,* but it was not until the modern period that Arabic numbers in the French *chevé* style (apparently learned in Germany by Tanaka Shōhei) were combined with Western rhythmic and measured devices to create notations that could be sight-read without the aid of a teacher. Three variations on this technique form the basis of most modern samisen notations, although occasional pieces can be found in Western notation as well. Thus, it is possible to purchase large repertoires of *nagauta, kouta,* or *kiyomoto* music for performance or study alone. Motivating such notational changes was the increased interest during the mid-19th century in samisen music composed for concert performance (*ozashiki*) rather than as dance accompaniment. Such a tradition is common practice for all the samisen genres today.

Offstage music. Returning to the theatre, one finds rather different music offstage. This *geza,* or *kagebayashi* (shadow *hayashi*), music is normally placed in a small room on stage right with a view of the drama through a bamboo curtain. The music consists of special samisen and vocal pieces and a great variety of percussion signals. For example, a huge *ō-daiko* barrel drum with two tacked heads signals the beginning of a program, in keeping with the sounds given by the same drum from a tower over the entrance of very early kabuki theatres. Other drums, bells, gongs, and clappers are used to reinforce stage action, and special offstage songs may set the mood or location of a scene, particularly in those scenes in which onstage musicians do not appear. For example, the singing of the offstage song "Eight Miles to Hakone" will tell an audience that the scene is set along the old Tōkaidō (the ancient road between Tokyo and Kyōto), whereas the sound of waves (*nami-no-oto*) beat on the *ō-daiko* drum indicates that the scene is on the road near the sea. A type of offstage song called *meriyasu* may be used to reflect the silent thoughts of the stage character, while the call and response of occasional beats offstage of two *ko-tsuzumi* drums will place a scene in a mountain area with its echoes. As in Western musical theatre and films, many of the sounds are naturalistic, whereas others are traditional means of evoking desired responses from an audience.

Music and dramatic content

Theatregoers in both traditions are often unaware consciously of the means used for such reactions even though familiarity has made their dramatic value very real indeed. The specific musical devices used in a given kabuki play are under the control of a headman, *hayashi gashira,* who works with the first samisenist, the actors, and the director to produce the desired results. Thus, the musical contents of a given play may change with different productions. In kabuki the combination of offstage and onstage music creates a total atmosphere that has few parallels in other world theatres. Perhaps it comes as close as anything to the composer Richard Wagner's ideal of the all-embracing art form (*Gesamtkunstwerk*).

Biwa, vocal, and folk music. During the late 19th century the *biwa*-accompanied narratives enjoyed a revival. The blind-priest *biwa* (*moso biwa*) tradition had originally been divided into two schools named after the provinces in Kyushu from which they came, Chikuzen and Satsuma. The tradition declined greatly over the years. When the Imperial restoration began in the Meiji period, many members of the new administration were from those provinces. Thus new schools of narrative *biwa* music arose under those two names, influenced at this time by several samisen narrative traditions. The topics of the new *biwa* pieces were often military and appropriate to the modernization period. The 19th century also was one of Japan's periodic revivals of interest in things Chinese, reduced somewhat with the advent of the Sino-Japanese War of 1894–95. Another late Tokugawa period style was *shigin,* the singing of Chinese poems in an intense solo style quite unrelated to the Heian *rōei* tradition of Chinese-based songs. *Shigin* was later accompanied by *shakuhachi,* and during the increased military spirit of the Meiji period it was combined with a posturing sword dance, *tsurugi-mai.* It also appeared in *biwa* concerts and could still be heard on rare occasions after World War II.

Work songs, theatricals, and festivals

Courtly writings have left little information about the music of the peasants in any detail, but some folk songs and theatricals of the Tokugawa period remain for modern study. The rice-planting, harvesting, and other work songs that survive may retain ancient melodies and may also be evidence of the indigenous origins of the *yo-in* scale systems to which most such music belongs. In this context, the first phrase of the folk song "Kuroda-bushi" is shown in notation X-D as it is said to have been derived originally from a Heian period *imayō* based on the gagaku piece shown in X-A. Most folk songs are, of course, regionally functional but historically vague and subject to the normal changes of any oral tradition. Viewing as a whole both the performance practice and voice qualities of Japanese folk music, one finds a great variety of styles. Such richness may reflect the long periods of Japanese feudalism, which fostered many different musical dialects.

Thatched many processionals and pantomimes of folk theatricals are accompanied by flutes and percussion, the generic term for such ensembles being *hayashi.* During the Tokugawa period the Shintō shrines of Edo (Tokyo) developed festival ensembles (*matsuri hayashi*) for the various major districts of the city. Most of these combine a bamboo flute with two folk-style *taiko* stick drums, an *ō-daiko* barrel drum, and a small hand gong called the *kane,* or *atarigane.* When such groups are playing general festival music, they all use a suite of five pieces: *yatai, shoden, kamakura, shichome,* and another *yatai.* However, their versions of each piece can be very different. When dance or pantomime is involved, the *sato-kagura* music mentioned earlier is used. The *kagura-bue* flute is often replaced by the nō flute. It combines with an *ō-daiko* and a *diabyoshi* barrel drum. The patterns on the heads of the latter contain East Asian male-female designs. One head is struck with thin bamboo sticks, the drum sitting to the side so that the player can better see the dancer. Lion dance (*shishi mai*) ensembles often use a trio consisting of a bamboo flutist, a gong player, and a drummer who plays a *taiko* and a small *odeko* barrel drum. Cymbals (*chappa*) and samisen may appear in other folk pantomimes or dances. The most common folk dances are the summer *bon odori,* traditionally performed in circles around a high platform (*yagura*) where the musicians or tape machines are located.

Given the oral base of all folk music, many songs are lost with the demise of another old farmer or worker. Scholarly and commercial interest in national music remains strong, however. Folk song preservation societies (*minyo hozon kai*) exist whose functions are to preserve "correct" performances of a single folk song. Such specificity seems unique to Japan. Regional and international folk-based Japanese ensembles flourish, and the summer dances can be seen in Japanese communities from Tokyo to Detroit.

Preservation of folk music

THE MEIJI PERIOD AND SUBSEQUENT MUSIC

Sources of Western influence. The period of Japanese history after 1868 is often thought of primarily in terms of its Westernization. The three major sources of Western music in Japan were the church, the schools, and the military.

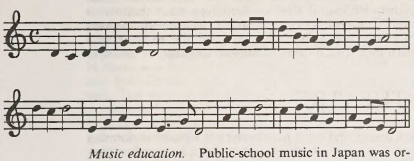
Religious and military music. Christian music had, in fact, been introduced into Japan as early as the mid-16th century with the arrival of Portuguese merchants and Roman Catholic priests. With this importation came Catholic music and Western musical instruments, the most lasting of which was the double-reed shawm, which survives today as the tuneful accessory of itinerant noodle sellers. The bowed *rebeca* lute may have combined with the Chinese *hu-ch'in* in the creation of the bowed *kokyū* of 17th-century Japan. However, the suppression of Christianity in that century destroyed the bamboo organs, choirs of mass singers, and most of the other direct Western musical imitations until the Meiji restoration. The official doctrine of new religious freedom in 1872 brought large numbers of Protestant missionaries into action, and collections of hymns with Japanese text were printed by 1878. Interdenominational editions were necessary by the 1890s. Since that time, standard Catholic and Protestant musical

activities can be found and, with the international growth of Tokyo, one can even add the sounds of synagogues and a mosque. But the growth of musical acculturation in Meiji Japan is better seen in its other foreign imports.

Band music, as part of a military table of organization, had already been tried in Dutch style at a military school in Nagasaki during the early 19th century. After Matthew C. Perry's arrival in 1853, every foreign delegation to Japan did its best to impress the natives with marching bands (Perry added a minstrel show). Thus, the various Japanese regional and national military leaders were quick to add such organizations to their modernized armies. The emperor was equally aware of the Western musical values displayed by the first foreign missions and ordered that the gagaku musicians be trained in band music as well. A navy band from the Satsuma clan gave the first Japanese public performance of this new music at the opening of the railroad in 1872, and in 1876 gagaku musicians made their debut as band musicians on the occasion of the emperor's birthday. The training of the many new ensembles was in the hands of English, French, and German bandmasters, and new music was created by them or by their Japanese students to match the spirit of Meiji modernism. The most famous case is the national anthem, "Kimi ga yo," which was one of the few successful early attempts at combining Western and Japanese traditions. A British bandmaster, William Fenton, teaching the Japanese navy band, worked together with gagaku musicians through several unsuccessful versions; and the search continued through his German successor, Franz Eckert. A court musician, Hayashi Hiromori (1831–96), is credited with the melody shown in notation XIV, which was given its premiere in 1880 and has remained the national anthem since that time. Hayashi first wrote it in traditional gagaku notation; and Eckert "corrected" it with Western harmonization, noting that it fit in both a gagaku mode (*ichikotsu*) and one from the Western church tradition (Dorian). As Japan's military prowess grew, standard Western-style marches and patriotic pieces dominated the repertoire. They also influenced popular music with such genres as *rappa-bushi* (literally, "bugle songs") as well as music in the schools.

Collaboration of Western and traditional Japanese musicians

XIV



Music education. Public-school music in Japan was organized by a member of a Meiji educational search team, Izawa Shūji (1851–1917), and a Boston music teacher, Luther Whiting Mason (1828–96). Mason was brought to Japan in 1880 to help form a music curriculum for public schools and start a teacher-training program. Although there was much talk of combining the best of East and West, the results of the sincere efforts of an American late-Victorian and a Japanese bureaucrat were less than glorious. The first children's songbook, the *Shōgaku shōkashū* (1881), contained either Western pieces with Japanese words or songs newly composed by Mason.

The primary sources of Western tunes were those pieces from Boston schoolbooks that appeared to be pentatonic. Through this method songs like "The Bluebells of Scotland" spoke of beauty ("Utsukushiki"), "Auld Lang Syne" concerned fireflies, and Stephen Foster became the major composer of songs known to educated Japanese children. The newly composed songs with their artificial tunes and moralistic words quickly faded away and eventually were replaced by more popular children's school songs based on military music (*gunka*) from the Sino- and Russo-Japanese wars. The teacher-training school became the Tokyo School of Music by 1890 and included instruction in koto and, because of the lack of proper violins, the bowed *kokyu*. The music department of the modern Tokyo University of Fine Arts and Music is still located at the spot of the original school in Ueno Park, Tokyo, with a bust of Beethoven beside the entrance. Koto, samisen, nō music, and Japanese music history are now found there, along with extensive offerings in Western music. However, until the late 20th century, music education was totally Western in orientation. Japanese music was presented in middle-school music appreciation courses only some 10 years after the end of World War II. The teaching of Western-style singing and the use of choruses have become fundamental to a proper education in Japan, with the results that youth and workers' choruses of the 20th century are cut off from original Japanese music. It was only with the rise all over the world in the mid-20th century of searches for cultural or ethnic identities that the Western nature of Japanese music education has been bypassed by some youth. Such a move should be quite clear to followers of Euro-American folk- and minority-group music revivals. In Japan, 20th-century activists, right or left, have attracted youth by the use of the public-school choral tradition in new textual contexts. Behind the robust volume of such functional, harmonized tunes lies the equally viable if quieter sounds of older, traditional music.

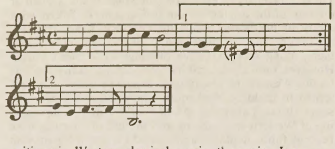
Traditional styles. The pre-Meiji period of 19th-century Japanese traditional music, known generically as *hōgaku* vis-à-vis Western music (*yōgaku*), was generally strong. It has been noted that certain styles of samisen music had been able to create concert repertoires disconnected from dance or party accompaniment. Koto teachers and composers also flourished; and *biwa* music began to return along with court music, paralleling the restoration of Imperial power. The most devastating effect of the restoration was the canceling of monopoly privileges previously held by the various guilds, including those in the music fields. This temporary economic-social setback was overcome by the admission of students from all classes of people and, at the same time, by a concerted effort on the part of more imaginative musicians to make some compromise between their old traditions and the new sounds flowing in from the West. In general, the evaluation of Western music by Japanese traditionalists showed that it differed from *hōgaku* in the following ways: it used other tone systems; it was thicker in texture, with more high and low notes going on at the same time; part of this thickness was sets of chords; it was generally considered better if it were faster and louder and the instruments were played more fancifully; it used more instruments at a time; it used different kinds of metres; and it had other forms, often organized by the concept of first and secondary themes. A survey of late 19th-century and early 20th-century musical experiments in Japan shows that every one of these characteristics was tried out, particularly in koto and samisen music.

Evaluation of Western music by Japanese traditionalists

Perhaps the most obvious and successful composer in the new traditional music (*shin hōgaku*) following World War I was Miyagi Michio (1894–1956), a blind koto teacher in the Ikuta school. In 1921 he composed a piece "Ochiba no odori" ("Dance of the Falling Leaves"), which used two koto, samisen, and a 17-stringed bass koto of his invention. Later works by Miyagi combine orchestras of traditional instruments, sometimes with strikingly successful results, although concerti for koto by some composers, with their mass koto and *shakuhachi* accompaniments, rather negate the entire sound ideal of the original idioms. The 1929 duet for *shakuhachi* and koto, "Haru no umi" ("Spring Sea"), has proven Baroque-like in its performance practice, for it is often heard played by the violin, with koto or piano accompaniment. Its style equals the French composer Claude Debussy in his most "orientale" moments. The Japanese traditionalist's view of Western music described above continued to be employed after World War II with such works as multimovement pieces using mixed orchestras in other contemporary idioms, including electronic manipulations. Such trends are best seen in the context of Western-style Japanese composers.

Composers in Western styles. Although graduates of the Tokyo School of Music and modernized court musicians were involved in many of the first concerts and com-

XV



positions in Western classical music, the major Japanese forces in this direction came from young men who studied in Europe. The most famous surviving composition of this era is *Kojo no tsuki* (*The Ruined Castle*), written in 1901 by Taki Rentarō after his training in Germany. The first line, shown in notation XV, reveals, with its use of E or E♯, a conflict between the Western minor and the Japanese *in* scales. In its piano-accompanied version it recalls the style of Franz Schubert, but as sung in the streets it sounds Japanese. Yamada Kōsaku was training in Germany when the Meiji era ended (1912) and returned to Japan with a new name, Koscak, and a strong interest in the founding of opera companies and symphony orchestras, as well as in the teaching of Western music. His opera, *Kurobune* (1940; *The Black Ships*), deals with the opening of Japan to the West and reflects his knowledge of Wagnerian style. Attempts at nationalistic operas can be represented better by the work *Yuzuru* (1952; *Twilight Crane*) by Ikuma Dan. The plot is a Japanese folktale, and, although the musical style is a mixture of the music of Maurice Ravel and the late works of Giacomo Puccini, one finds as well deliberate uses of folk songs and idioms. Shimizu Osamu is perhaps more successful nationalistically in his choral settings of Japanese and Ainu music, in which the style of vocal production and chordal references seems to be a more honest abstraction of Japanese ideals. Mamiya Michio combined traditional timbres with 12-tone compositional technique in a koto quartet. Mayuzumi Toshirō has produced many clever eclectic results in such works as his *Nirvana Symphony* (1958); Buddhist sutra texts mix with a combination of choral writing in the style of Igor Stravinsky, orchestral tone clusters, and sweeping vocal lines derived from Japanese Buddhist chant style.

It has often been felt that no true combination of Japanese and Western music would be possible until there was some composer who was equally knowledgeable in both Western and Japanese traditional styles. Such a musical, aesthetic barricade seemed unbroken until the last third of the 20th century, when international music styles made culturally transcendental eclecticism a viable medium for those composers with enough talent and insight to control the infinite idioms available to them. In Japan, Takemitsu Toru seems a likely candidate for such an accolade. His music is totally contemporary and never directly "orientale," yet some of his senses of timing, texture, and structure are characteristically Japanese.

20th-century musical internationalism

In modern Japan all styles of music are available, from the traditional to the most avant-garde. Fully professional performances of kabuki music are matched by complete Beethoven symphonic series. Huge choruses singing polemics of every type and mass bands of children bowing violins in the widely imitated method of instruction developed by Shinichi Suzuki compete for audiences with intimate recitals of Heike *biwa* music and hundreds of other events. Research in Japanese traditional music has flourished among native scholars as well as among an increasing number of foreign devotees; and national, private, and academic organizations have been founded for the collection, study, and publication of material dealing with all aspects of Japanese musical life.

From the outline of Japanese musical culture given above, it should be evident that old traditions can still be heard along with the newer ones. For the most part, the older forms probably do not sound the same today as they did in their heyday. Such changes in traditions are inevitable, however, and are common to music in most other world cultures, including the Western. For example, present-day gagaku performances are undoubtedly different from those of 1,000 years ago, but Mozart symphonies as well do not sound the same as they did in the 18th century. Now modern technology has made it possible to "freeze" a given performance of some musical event through a recording. Each musician in each generation may choose as he desires to add fresh flavour to such earlier items or leave them "pure." Part of the charm and fascination of Japanese music is that it still offers so many stylistic listening and studying choices to anyone curious or energetic enough to want to know them better. A major point of this entire discussion is that none of the various styles of East Asian music is any more mystical or incomprehensible than is Bach or Beethoven. Each tradition is simply different. All of them are also logical and—perhaps of greater importance—they are beautiful to those who learn their special forms of musical language. (W.P.M.)

DANCE AND THEATRE

Origin of dance and theatre in religious customs

From ancient times dance and theatre have played a vital role in China, Korea, and Japan. Many performances of plays and dances were closely tied to religious beliefs and customs. In China, records from about 1000 BC describe magnificently costumed male and female shamans who sang and danced to musical accompaniment, drawing the heavenly spirits down to earth through their performance. Impersonation of other characters through makeup and costume was occurring at least by the 4th century BC. Many masked dances in Korea have a religious function. Performances invoking Buddha's protection are especially popular and numerous in Japan and Korea. Throughout East Asia the descendants of magico-religious performances can be seen in a variety of guises. Whether designed to pray for longevity or for a rich harvest or to ward off disease and evil, the rituals of impersonation of supernatural beings through masks and costumes and the repetition of rhythmic music and patterns of movement perform the function of linking man to the spiritual world beyond. Hence, from the earliest times in East Asia, dance, music, and dramatic mimesis have been naturally fused through their religious function.

In East Asia the easy intermingling of dance and theatre, with music as a necessary and inseparable accompanying art, also derives from aesthetic and philosophic principles (see above). In the West, by contrast, concert music, spoken drama, and ballet have evolved as separate performing arts. Confucian philosophy holds that a harmonious condition in society can be produced by the proper actions of man. Throughout China's history, poems were written to be sung; songs were danced. Zeami (1363–1443), the most influential performer and theoretician of nō drama in Japan, described his art as a totality, encompassing mimesis, dance, dialogue, narration, music, staging, and the reactions of the audience as well. Without arbitrary divisions separating the arts, there has developed in East Asia exceptionally complex artistic forms that produce on their audience an impact of extraordinary richness and subtlety.

Puppets, masks, highly stylized makeup, and costuming are common adjuncts of both dance and theatre. Dialogue drama (without music) is rare but does exist. The major dance and theatre forms performed today in East Asia can be loosely classed as unmasked dances (folk and art dances in each country), masked dances (Korean masked dances and *bugaku* and folk dances in Japan), masked dance theatre (nō in Japan and *sandae* in Korea), danced processionals (*gyōdō* in Japan), dance opera (Peking and other forms of Chinese opera), puppet theatre (*kkoktukaksi* in Korea and bunraku in Japan), shadow theatre (in China only), dialogue plays with traditional music and dance (kabuki in Japan), dialogue plays with dance (*kyōgen* in

Current dance and theatre forms

Japan), and modern, realistic dialogue plays introduced from the West into China, Korea, and Japan in the 19th and 20th centuries.

Characteristics of East Asian dance and theatre

COMMON TRADITIONS

As previously noted, China, Korea, and Japan have been historically close for centuries, thus accounting for their numerous common artistic traditions. From pre-Christian times until the 8th and 9th century AD, the great trade routes crossed from the Middle East through Central Asia into China. Hinduism, Buddhism, some knowledge of ancient Greek, and much knowledge of Indian arts entered into China, and thence in time into Korea and Japan. Perhaps before Christ, the Central Asian art of manipulating hand puppets was carried to China. For more than 700 years, until 668, in the kingdom of Koguryŏ, embracing northern Korea and Manchuria, court music and dances from Central Asia, from Han China, from Manchuria, and from Korea, called *chisŏ* and *kajisŏ*, were performed. Many of the dances were masked; all were stately as befit serious court art. They were taken to the Japanese court in Nara about the 7th century. Called *bugaku* in Japan, they have been preserved for 12 centuries and can still be seen performed at the Imperial Palace in Tokyo, though they have long since died out in China and Korea. In Koguryŏ's neighbouring kingdom of Paekche, a form of Buddhist masked dance play was performed at court, and, in the 7th century, it too was taken to the Japanese court at Nara by a Korean performer, Mimaji, who had learned the dances while staying at the southern Chinese court of Wu-hou. Called *kiak* in Korea and *gigaku* in Japan, the Aryan features of some of its masks clearly indicate Indian (or Central Asian) influence. Such complicated genealogies are common in East Asian performing arts.

Very likely by the 7th century, gypsylike puppeteers, who originally had been nomads from Central Asia and had taken up abode in northern Korea, migrated to Japan. (There may have been a native puppet tradition in Japan as well.) In time the art of puppet manipulation joined with that of epic storytelling to produce the famous bunraku puppet theatre. Musical accompaniment for bunraku and for other popular plays in Japan, such as kabuki, was provided primarily by the samisen, a three-stringed lute, borrowed from China by way of Okinawa. The lion dance, originally from China, is performed in a score of versions in Korea and Japan as well as in China, India, Sri Lanka, and Bali. Certain myths are dramatized in common as well. The story of the angel or nymph who flies down to earth and arouses the love of a mortal man is known in many parts of the world. It is dramatized in Southeast Asia (especially in Myanmar [Burma] and Thailand, as the play *Manora*), in Chinese opera, and in both nō and kabuki theatre in Japan (as *Hagoromo* [*The Feather Robe*]). The legend of the one-horned wizard who traps the dragon gods of rain and causes a searing drought originated in India and was later transmitted by the Chinese to Japan, where it is dramatized in nō (*Ikkaku sennin* ["The One-Horned Wizard"]) and in kabuki (*Narukami* [*Saint Narukami and the God Fudō*]).

Western influences

The direction of artistic exchange was reversed in the 19th century. As part of Japanese national policy following the Meiji Restoration (1868), artists studied Western performing arts. In the early decades of the 20th century, Chinese and Korean actors, dancers, and playwrights studying in Japan took back to their countries Western theory and practice in ballet, modern dance, and theatre. Most influential was the Western dramatic theory of realism. It diametrically opposed the traditional intermingling of music and dance with drama, and it eschewed the stylization and symbolism that lay at the heart of East Asian performing arts for more than 2,000 years. A conflict between traditional and Western performing arts came into being that continues to the present.

As has been noted, dance and theatre are accompanied by music in all except the most unusual cases. The music is especially composed for each bunraku puppet play in Japan and for most dance plays and court dances. Fixed melodies accompany most folk performances. In Chinese opera and in Japanese kabuki, melodies appropriate to scene, action, character, or mood being portrayed are selected from a standard musical repertoire of several hundred tunes. The knowledgeable spectator easily identifies scenes by the music that accompanies them (a similar system is found in Southeast Asian theatre). The close linking of music with dance and theatre can be seen in the Korean drum dance, in which the dancer also is a musician who plays the drum, and in a number of Japanese kabuki and puppet plays that show characters expressing hidden feelings by playing a musical instrument. Equally important, the performer demonstrates to the audience his skill in yet another refined accomplishment.

The performing arts of India are closely linked to sculpture and painting by the unusual phenomenon that bodily positions in all these arts are regulated by similar, indeed almost identical, codes. The code of hand gestures, for example, for the dancer and the actor set forth in the *Nāṭyaśāstra* ("Treatise on the Dramatic Arts"; dated variously from the 2nd century BC to the 4th century AD and later), a treatise on dramaturgy, is identical with that for Buddhist temple sculpture, or painting. Although these hand positions (*mudrā*) from India also are seen in Chinese, Korean, and Japanese statues of the Buddha, they have never been adopted by performing artists (as, by way of contrast, they were by dancers and actors in Cambodia, Thailand, Myanmar, Java, and Bali).

Aesthetic parallels of visual arts and performing arts

In three notable instances, however, the performing arts in China and in Japan can be seen to be closely related to the visual arts. During the Sung dynasty (960–1279) in China, Northern and Southern schools of painting evolved that were totally different in style; the former used bold outline and brilliantly contrasting colours of deep green, blue, and gold, while the latter emphasized delicate, monochrome ink painting of misty landscapes. Northern and Southern schools of opera at the time reflected the same contrasting characteristics: the former dynamic, vigorous, and filled with action, the latter emphasizing wistful emotions and soft, gentle singing. Zen Buddhism was a common source of inspiration in the 15th and 16th centuries in Japan for nō dance drama, for the tea ceremony, for ink painting, and for the art of rock-and-sand gardens. Spareness of form, discipline, and suggestion rather than explicit statement are Zen attributes found in these and other arts cultivated by the military ruling class (samurai) of the time. In 18th-century Japan, a lively and faddish urban culture produced both ukiyo-e woodblock prints and kabuki. In fact, ukiyo-e artists, such as Tōshūsai Sharaku, established their fame by portraying famous kabuki actors as their subjects. Eroticism, verve, brilliant colouring, and an intense interest in the passing moment characterize equally both kabuki theatre and ukiyo-e visual art.

The performing arts traditionally are seen as distinct from literature in East Asia. A century and a half passed in kabuki before the first complete play script was preserved, and in China, where a tradition of written literature goes back to 1400 BC, no play text was considered worth committing to paper until the late Sung, about the 13th century. With few exceptions, playwrights have rarely been accorded the same status as writers of poetry, novels, or criticism. As a result, the performing arts in East Asia succeeded by and large in escaping the stultifying grip that literature came to hold on Indian Sanskrit drama and, some would say, still holds on drama in the West.

Local traditions in East Asian countries

The general outlines of artistic borrowings among East Asian countries can be traced from historical records. But borrowing tells only half of the story. No matter how strong the initial outside influence, in time, assimilation of the foreign art took place. Older native performing traditions reasserted themselves, and new creativity altered the borrowed elements. This can be seen even in *bugaku* dances in Japan; although they are believed to preserve ancient Chinese and Korean forms to a very remarkable extent, native Japanese qualities are also present. Local styles predominate even more in the popular arts. Japanese bunraku puppet plays and kabuki theatre show almost no observable signs of foreign influence. In spite of certain

general cultural similarities, then, the dance and theatre of China, Korea, or Japan exhibit definite local characteristics not shared by the arts of their neighbouring countries.

In China singing became highly developed, and the most important theatre performances are built around song (hence the term Chinese opera). The shadow theatre, known from Morocco through Egypt and Greece and in India, Indonesia, Malaysia, Thailand, and Cambodia, is found in East Asia only in China. In Korea there are scores of court and folk dances and danced plays, but no sophisticated dramatic forms evolved until the 20th century. Masked dances especially are characteristic of Korea. In Japan complex theatrical forms evolved that include dance drama, epic narrative performed as a puppet play, and dialogue dramas either accompanied or unaccompanied by music.

Aesthetic principles

The aesthetic principles that govern dance and theatre in East Asia are radically different from those of the West. Dancers in the West attempt to be free from the pull of the Earth, trying to leap and soar in the air. Dancers in China, Korea, or Japan stand firmly on the dance floor, often scarcely raising their feet in the air; they move in relatively slow and often geometric patterns. Arm and hand movements are important and varied, while in Western dance the hands are little used. Whether movement is dance or not, it is always stylized. Speech is stylized as well, whether it is dialogue or narration, chanted or sung. The intent may be to portray archetypes, human or mythological, especially in shadow and puppet theatre and in masked dances and plays. There is great emphasis on form, both for its ritual value and because audiences are trained to recognize the beauty implicit in form. The East Asian audience is prepared to respond in quick succession to a sequence of different stimuli—physical characterization, human speech, song, narrative commentary, visual composition, formal movement patterns—over long periods of time, for 8 hours in nō or up to 12 hours in kabuki. This differs from the West, where the spectator expects to be exposed to a clearly focused theatrical image for only two or three hours. The East Asian experience is more diverse, more extended, more conventional than the Western experience in the theatre.

A further important characteristic of dance and theatre in China, Korea, and Japan is that performing arts developed very largely within an oral tradition. By and large the performers themselves created the forms; only gradually did specialists in choreography, musical composition, or writing take their places in performing groups.

SOCIAL CONDITIONS

Social position of the performer

It is notable that, although some dance and theatre forms were highly regarded in China, Korea, and Japan, performers were usually looked down upon. Wandering performers especially were despised in the agrarian societies, where attachment to the land was valued and Confucian teaching, strong throughout East Asia, stressed veneration of one's parents, which included tending their graves and making offerings for their welfare in the spirit world. The place of drama or of dance in these societies depended in part upon their audiences, whether they were court nobles, villagers, or town merchants.

Chinese emperors, Korean kings, and Japanese emperors and military rulers (shoguns) all supported performers at their courts. During the T'ang dynasty, the 8th-century Chinese emperor Hsüan-tsung (also called Ming-huang) established schools in the palace city of Ch'ang-an (Sian) for music, dancing, and acting. The latter school was called the Pear Garden (Li-yüan); ever since, actors in China have been called "children of the pear garden" (*li-yüan tzu-ti*). More than a thousand young people from all ranks of society drew government salaries while studying and performing at lavish state banquets and for official ceremonies. Acting or dancing might be a permanent job (at least until old age made one less attractive) at the Chinese court, but in Korea performers at the court held other positions in the government and were mobilized from around the country only for rehearsal and performance. In Japan, dancers and musicians have been attached to the Imperial household from the 7th century until the present time. First *gigaku* and then *bugaku* dances were official performing arts, while shrine dances (*kagura*) were also partly under Imperial patronage. The military rulers of Japan incorporated into their retinues nō actors and musicians beginning in the 15th century, and, in time, provincial lords also began to follow this practice.

Court patronage

Court support resulted in high artistic levels in all countries. Performers were relieved of financial problems and could devote themselves, often full-time through their entire lives, to their art. Audiences were educated and for the most part discerning. The importance attached to official performances undoubtedly spurred artists to extend themselves to their utmost. In time, however, such forms as Japanese nō and *bugaku* and Chinese *k'un-ch'ü* opera became so rarefied that they could be appreciated only by a small elite group.

At the Chinese and Korean courts, young female dancers were part of the ruler's personal retinue (often his concubines); they were not allowed to mix with men of the court, so that some court arts were performed solely by men and others solely by women. This custom and the consequent artistic practice of male and female impersonation is also found in court theatre of Cambodia and Thailand. In Japan, women seldom performed at court, and the major dance and theatre forms have been the province of male performers. Since it was unusual for rulers or courtiers themselves to take part in performance, the court artist was usually a middle-level civil servant.

Folk and popular theatre and dance

Folk performers, on the other hand, are local villagers who, like the *sandae* masked dancers of Korea or the young women who perform festive *ayakomai* dances in Japan, are amateurs who do not live by their art. The midsummer Bon dance for spirits of the dead or early spring rice-planting dances in many areas of Japan or various auspicious dances held at the New Year in Korea and China were performed only once a year, and hence a high level of artistry was not usually achieved. Because many folk performances were held as part of religiously sanctioned rituals (Korean mask plays ensuring harvest, dances and dance plays of many varieties in Japan dedicated to local Shintō deities), performers achieved considerable status in the local community by their participation in these essential communal rites.

Performers of popular dance and theatre in East Asia live—as do commercial artists everywhere—by their ability to draw audiences who are willing to pay money for a seat in a public theatre. The shadow and puppet performers of China, Peking opera actors and musicians, and kabuki and bunraku puppet performers in Japan are popular artists. Neither a part of village culture nor patronized by the court, they have always been held suspect by their rulers. Popular theatre grew in importance in China and in Japan concurrently with the growth of large urban centres and a moneyed, mercantile economy, in the 17th–19th centuries. Today troupes perform nightly through the year, when it is possible, and consequently, in popular theatre, large repertories of standard plays are created (some 350 in kabuki and more than 200 in Chinese opera). Popular theatre forms in China and Japan are intensely theatrical, though they lack literary qualities which would recommend them to the intelligentsia. Indeed, kabuki in Japan and Peking opera in China have had little official status until the mid-20th century in spite of their immense audience popularity and their obvious excellence as performing arts.

The development of dance and theatre in the East Asian nations

CHINA

Formative period. Singing and dancing were performed at the Chinese court as early as the Chou dynasty (*c.* 1111–255 BC). An anecdote describes a case of realistic acting in 402 BC, when the chief jester of the court impersonated mannerisms of a recently deceased prime minister so faithfully that the emperor was convinced the minister had been restored to life. Drama was not yet developed, but large-scale masques (a short allegorical performance with masked players) in which dancing maidens and young

boys dressed as gods and as various animals were popular. Sword-swallowing, fire-eating, juggling, acrobatics, ropewalking, tumbling, and similar stage tricks had come from the nomads of Central Asia by the 2nd century BC and were called the "hundred entertainments." During the Han dynasty (206 BC–AD 220) palace singers acted out warriors' stories, the forerunners of military plays in later Chinese opera, and by the time of the Three Kingdoms (AD 220–280) clay puppets were used to enact plays. These evolved into glove-and-stick puppets in later years.

T'ang period. The emperor Hsüan-tsung showed interest in the performing arts, stimulating many advances in stage arts during the T'ang dynasty (618–907). More than a thousand pupils were enrolled in music, dance, and acting schools. Spectacular masked court dances and masked Buddhist dance processions were part of court life. Three types of play are recorded as having been popular. *Tai-mien* ("Mask") was about Prince Lan Ling, who covered his gentle face with a horrifying mask to frighten his enemies when he went into battle. Some suggest the colourful painted faces of warriors in today's Chinese opera derive from this play. *T'a-yao niang* ("Stepping and Swaying Woman") was a farcical domestic play in which a sobbing wife bitterly complained about her brutal husband, who then appeared and, singing and dancing, abused his wife even more. The embezzling rascal hero of *Ts'an-chün* ("The Military Counselor") became a stock character in later plays. Thus, by T'ang times, three basic types of drama were known: military play, domestic play, and satire of officialdom; and establishment of role types had begun.

Popular types of plays

Sung period. The variety play (*tsa-chü*) was created by writers and performers in North China during the Northern Sung dynasty (960–1127). None of the scripts has survived, but something of their nature can be deduced from the 280 titles which remain and from court records. A play consisted of three parts: a low-comedy prologue, the main play in one or two scenes (consisting of extended sequences of songs, dancing, and perhaps dialogue), and a musical epilogue. Two, three, or four variety plays would be included in a program along with a sampling from the "hundred entertainments." In the following Southern Sung dynasty (1127–1279), northern writers continued composing plays of this general type. None of the 691 professional scripts of which the titles are known has survived. Concurrently a new form of drama, southern drama (*nan-hsi*), emerged in the area around Hang-chou in southern China. Originally the creation of folk authors, it soon became an appealing and polished dramatic form. A southern drama tells a sustained story in colloquial language; flexible verses (*ch'ü*) were set to popular music, making both music and poetry accessible to the ordinary spectator. Professional playwrights belonging to Hang-chou's writing societies (*shu-hui*) wrote large numbers of southern dramas for local troupes. Of these, 113 titles and 3 play texts remain, preserved in an imperial collection of the 15th century. *Chang Hsieh chuang-yüan* ("Top Graduate Chang Hsieh") is probably the oldest of the three texts. It dramatizes the story of a young student who aspires to success, earns a degree and position, but callously turns his back on the girl who faithfully loves him.

Professional theatre districts became established during the Sung dynasty. Major cities contained several districts (17 or more in Hang-chou), with as many as 50 playhouses in a district. Plays performed by puppets and mechanical dolls were extremely popular.

Shadow theatre

A legend attributes the origin of shadow theatre in China to an incident said to have occurred about 100 BC: a priest, claiming to have brought to life the emperor's deceased wife, cast a woman's shadow on a white screen with a lamp. Others suggest the shadow play dates only from the Sung period. In any case it was widely performed in Sung times in the theatre districts. Puppets were made of translucent leather and coloured with transparent dye so they cast coloured shadows on the screen. Shadow plays are still performed in China. Singers, dancers, actors, acrobats, and other performers were all employed at the professional theatres of the districts. Troupes were as small as possible for economic reasons, containing as few as five or six performers. They would tour the countryside if they had no work in the large cities, thus spreading urban styles of performing arts throughout the vast region of China.



Mounted general wearing armour and commander's headdress, Chinese shadow puppet; in a private collection.
Collection of Professor and Mrs. Derk Bodde. Center of Puppetry Arts, Atlanta, Ga., photograph, Mary Carolyn Pendar

Yüan period. Scholars turned to writing drama in the Yüan period (1206–1368) when they were removed from their positions in the government by China's new Mongol rulers, descendants of Genghis Khan. They developed the earlier northern style of *tsa-chü* into a four-act dramatic form, in which songs (in the same mode in one act) alternated with dialogue. Singing was restricted to a single character in each play. Melodies were those of the Peking region. The beauty of poetic lyrics was highly valued, while plot incidents were of lesser importance. About 200 plays survive, from the thousands of romances, religious plays, histories, and domestic, bandit, and lawsuit plays that were composed. *Hsi-hsiang chi* (*The Romance of the Western Chamber*), by Wang Shih-fu, is a 13th-century adaptation of an epic romance of the 12th century. The student Chang and his beautiful sweetheart Ying Ying are models of the tender and melancholy young lovers who figure prominently in Chinese drama. Loyalty is the theme of the history play *Chao-shih ku-erh* (*The Orphan of Chao*), written in the second half of the 13th century. In it the hero sacrifices his son to save the life of young Chao so that Chao can later avenge the death of his family. *Hui-lan chi* (*The Chalk Circle*), demonstrating the cleverness of a famous judge, Pao, is known in the West, having been adapted (1948) by the German playwright Bertolt Brecht in *The Caucasian Chalk Circle.* The class of bandit dramas are mostly based on the novel *Shui-hu chuan* (*The Water Margin*) and its 108 bandit heroes, who live by their wits doing constant battle against corrupt and avaricious officials. The life of the common man is portrayed with considerable reality in Yüan drama, though within a highly formalized artistic frame. The lasting worth of Yüan plays is attested to by the fact that they have been adapted constantly to new musical styles over the years so that Yüan masterpieces make up a large part of the traditional opera repertory.

Ming period. Plays of the Yüan period were widely popular with the people. When under the native Chinese Ming rulers (1368–1644) Mongol influence was eradicated, drama was, for a time, forbidden. Revived in the south, it increasingly became a literary form for a scholarly elite. A renowned Ming play is *P'i-p'a chi* ("Lute Song"), written in 42 affecting scenes, by the scholar Kao Ming in the 14th century. Its heroine, Chao Wu-niang, sets a perfect example of Confucian filial piety and marital fidelity, car-

ing for her husband's parents until their tragic death and then playing the lute to eke out a living as she patiently searches for her husband.

In the mid-16th century, a musician, Wei Liang-fu, of Su-chou, devoted 10 years to creating a new style of music called *k'un-ch'ü*, based on southern folk and popular melodies. At first it was used in short plays. Liang Ch'en-yü, poet of the 16th century, adapted it to full-length opera in time, and it quickly spread to all parts of China, where it held the stage until the advent of Peking opera, two centuries later. Important *k'un-ch'ü* dramatists were T'ang Hsien-tsu (d. 1616), famed for the delicate sensitivity of his poetry, Shen Ching (d. 1610), who excelled in versification, and the creator of effective theatrical pieces, Li Yü (1611–1685). A large-scale performance of *k'un-ch'ü* for the Ch'ing emperor Ch'ien-lung in 1784 marked its high point in Chinese culture. *K'un-ch'ü* had begun as a genuinely popular opera form; it was welcomed by audiences in Peking in the 1600s, but within decades it had become a theatre of the literati, its poetic forms too esoteric and its music too refined for the common audience. In 1853 Su-chou was captured by the Taiping rebels, and thereafter *k'un-ch'ü* was without a strong base of support and declined rapidly.

Origin of Peking opera

Ch'ing (Manchu) period. *Ching-hsi* or *ching-chü* (Peking opera) came into being over a period of several decades at the end of the 18th century, during the Ch'ing dynasty (1644–1911/12). In the wake of the Taiping Rebellion, *k'un-ch'ü* troupes resident in Peking returned to their homes in the south. Their places in Peking's theatres were quickly taken by opera troupes from the surrounding provinces, especially Anhwei, Hupeh, Kansu, and Shansi. Anhwei opera had been performed on the occasion of the emperor Ch'ien-lung's birthday in 1790. Peking opera was born of an amalgamation of elements from several sources: rhythmic beating of clappers to mark time for movements (from Shansi and Kansu), singing in the two modes of *hsi-p'i* and *erh-huang* (from Anhwei), and increased use of acrobatics in fighting scenes. Undoubtedly, court support for Peking opera from Tz'u-hsi (1835–1908), the Empress Dowager, contributed to its rise, but it was also very widely patronized by local audiences. It became the custom to rehearse in public teahouses, and in time these became regular performances providing troupes with much of their financial support.

Differences between *ching-hsi* and *k'un-ch'ü* drama

Essentially, *ching-hsi* was a continuation of northern-style drama, while *k'un-ch'ü* marked the culmination of southern-style drama. Musically they are very different: the former uses loud clappers and cymbals for scenes of action and the penetrating sound of fiddles accompanies singing; in the latter the flute is the major instrument, and strings and cymbals are absent. A limited number of melodies are repeated many times in Peking opera (set to different lyrics), while in *k'un-ch'ü* the melodic range is much wider. Peking opera lyrics are in colloquial language (they are often criticized as lacking in literary merit). Overall, the newer opera form is highly theatrical and vigorous, while the older form is restrained, gentle, and elegant. Some Peking operas are Yüan plays or *k'un-ch'ü* operas adapted to the new northern musical system. Many plays first staged as Peking opera are dramatizations of the war novel *San-kuo chih yen-i* (*Romance of the Three Kingdoms*), written in the 14th century by Lo Kuan-chung. Mei Lanfang, the most famous performer of *ching-hsi* female roles in the 20th century, introduced a number of these highly active military plays into the repertoire. *K'un-ch'ü* dramas told a long and involved story in great detail, often in 40 or 50 consecutive scenes. It became the custom in Peking opera to perform a bill of a number of acts or scenes from several plays, like a Western concert program.

© Wu Gang/Liaison International



A *ching-hsi* troupe performing a scene from *Pai-she chuan* (*The White Snake*).

Concurrent with the national forms of drama mentioned before, local opera is found in every area of China (the different forms have been estimated at 300). These operas are performed according to local musical styles and in regional languages. General characteristics of most forms of Chinese opera are similar, however. Action occurs on a stage bare of scenery except for a backdrop and side-pieces. A table and several chairs indicate a throne, wall, mountain, or other location. (More elaborate scenery is used in Canton and Shanghai, influenced by Western drama and motion pictures.) Actors enter through a door right and exit through a door left. Costumes, headgear, and makeup identify standard character types. Actors play a single role type as a rule: male (*sheng*), female (*tan*), painted-face warrior (*ching*), or clown (*ch'ou*). Each role type can be subdivided into several role subtypes. Actors undergo seven years of training as children, during which time their appropriate role type is determined. Singing is essential for *sheng* and *tan* roles; minor actors and actors of clown roles must be skilled in acrobatics that enliven battle scenes. Singing is accompanied by a large number of conventionalized movements and gestures. For example, the long "flowing water" sleeves that are attached to the costumes of dignified characters can be manipulated in 107 movements. Pantomime is highly developed, and several scenes have become famous for being enacted without dialogue: in *Pai-she chuan* (*The White Snake*) a boatman rows his lovely daughter across a swirling river; in *San cha kou* ("Where Three Roads Meet") two men duel in the dark; in *Shi yu chuo* ("Picking Up the Jade Bracelet") a maiden threads an imaginary needle and sews. Symbolism is highly developed. Walking in a circle indicates a journey. Circling the stage while holding a horizontal whip suggests riding a horse. Riding in a carriage is represented by a stage assistant holding flags painted with a wheel design on either side of the actor. Four banners indicate an army. A black flag whisked across the stage means a storm, a light blue one a breeze or the ocean. Chinese opera is one of the most conventionalized forms of theatre in the world. It has been suggested that the poverty of troupes and the need to travel with few properties and little scenery led to the development of many of these conventions.

Role of Confucian morality in drama

Confucian morality underlies traditional Chinese drama. Duty to parents and husband and loyalty to one's master and elder brother or sister were virtues inculcated in play after play. Spiritualism and magic powers, derived from Taoism, are themes of some dramas, but by and large Chinese drama is ethical rather than religious in direction. Plays were intended to uphold virtuous conduct and to point out the dire consequences of evil. The Western tragic view, which holds that man cannot understand or control the unseen forces of the universe, has no place in Chinese drama; the typical play concludes on a note of poetic justice with virtue rewarded and evil punished, thus showing the proper way of human conduct in a social world.

20th century. With the establishment of the Republic of China in 1912, court support for Peking opera by the Manchu dynasty ended. Troupes, however, continued to perform for private patrons and in public at teahouses and in theatres. Following the liberal ideals of the time, attempts were made to write in colloquial language (rather

than in classical Chinese, as previously), and old plays considered undemocratic were dropped from the repertoire. A school for Peking opera acting, modeled on Western pedagogical methods, was established in 1930, actresses being admitted for the first time in three centuries. The basic style of opera remained unchanged, however.

Influence of Western drama

Western spoken drama (*hua-chü*) was first introduced by Chinese students who had studied in Japan and there learned of Western plays. In 1907 a Chinese adaptation of *Uncle Tom's Cabin* was successfully staged in Shanghai by students, marking the beginning of a proliferation of amateur study groups devoted to reading and staging Western plays. Originally aimed at only a small group of Western-educated intelligentsia, spoken drama's appeal was broadened to the middle class by the China Traveling Dramatic Troupe, which toured many cities from its home in Shanghai. In 1936 it performed *Leiyu* (*Lei-yü; Thunderstorm*), a four-act tragedy by Cao Yu. An extremely successful playwright in the Western style, by 1941 Cao had written six important plays, including *Beijingren* (1940; *Pei-ching jen;* "Peking Man"); heavily influenced by Eugene O'Neill and Henrik Ibsen, he portrayed dissolute members of the old gentry class and new rising entrepreneur class.

Nationalism, the upheaval of World War II, and changes of government in China, Korea, and Japan between 1945 and 1949 are reflected in contemporary theatre and dance in East Asia. In China an estimated 60,000 performers were mobilized into some 2,500 propaganda troupes during the Sino-Japanese War beginning in 1937 under the direction of the well-known playwright Tian Han. Hundreds of thousands of ordinary Chinese in the army were exposed to modern forms of drama for the first time, and, equally significant, artists discovered regional folk legends, songs, and dances, which they then incorporated into their work. For example, *Baimao nü* (*Pai-mao nü; The White-Haired Girl*) was developed from northern Chinese *yang-ko* folk dances into both a ballet and an opera. The heroine, an escaped concubine of a cruel landlord, symbolized all victims of feudal governments and oppressive social systems.

At Yen-an in 1942 Mao Zedong enunciated one of the basic principles of communist art: art should have the dual function of serving the masses and of being artistically superior. In the years since the establishment of the People's Republic of China in 1949, theatre activities have swung between these goals, depending on the current ideological line of the government. Initially, the traditional opera repertoire was purged of feudal, superstitious, or otherwise ideologically incorrect material. Government policy encouraged realistic spoken drama (*hua-chü*); but, in spite of successes such as Lao She's naturalistic *Chaguan* (1957; *Ch'a-kuan;* "Teahouse"), audiences have not responded to this "foreign" form of drama. From 1964, when Jiang Qing, Mao's wife, guided the composition of the first modern revolutionary operas, in which contemporary soldiers and workers were the heroes, until 1977, traditional operas were completely banned. During the Great Proletarian Cultural Revolution (1966–76), many traditional theatre artists were denounced or imprisoned. Famous modern drama figures such as Wu Han, author of *Hai Rui baguan* (1960; *Hai Jui pa-kuan; Hai Jui Dismissed from Office*), were persecuted and their plays banned. With the fall of the Gang of Four in 1976, the traditional repertoire was reinstated once more and Jiang's "model" revolutionary operas no longer staged. During the decade-long open-door policy (1979–89), theatre contacts with the West were tentatively resumed after 40 years abeyance: Arthur Miller was invited to direct *Death of a Salesman* in 1983, and the Shanghai Kun-ch'ü Opera Company toured in Europe with its opera version of *Macbeth* in 1987. The influence of Western plays is seen in the social satire *Chia-ju wo shih chen-ti* (1979; "If I Were Real") by Sha Yexin and Gao Xingian's Artaudian *Ye ren* ("Wild Man"), initially banned, then produced in 1985.

Government policies strongly affect the economics of Chinese theatre as well as dramatic themes and forms. After the establishment of the People's Republic, professional theatre troupes received full government subsidy. Following economic liberalization policies of 1986–87, however, troupes were required to earn increasing revenues from box-office income. At the same time, urban audience attendance declined (in part because of competition from films and television), with the result that some troupes disbanded and others were reduced in size. Government-supported theatre academies in Peking, Shanghai, and regional capitals play an essential role in training young theatre artists in traditional as well as modern genres. Foreign theatre exchanges of the 1980s were welcomed by many theatre artists who wished to bring new ideas into Chinese theatre, in particular to appeal to youthful audiences who were abandoning theatre for film and television; these exchanges again were halted in 1989 in the wake of the government's suppression of the Chinese student democracy movement at Tiananmen Square.

The performing arts in Nationalist China

The Nationalist government has supported Peking opera on Taiwan since establishing the headquarters of the Republic of China on that island. Troupes of the air force and the army are active, and the Foo Hsing Opera School receives government support. Local opera (*kotsai-hsi*), sung in the Taiwanese dialect, is extremely popular in commercial theatres, and many itinerant Taiwanese troupes tour glove-puppet plays (*po-the-hi*) to towns and villages.

KOREA

In addition to folk dances, the main traditional forms that developed in Korea are ritual court dances, masked dances, and puppet plays. Of these, masked dances and masked-dance plays have perhaps the oldest and richest traditions. Archaeological evidence suggests that masks were used at least by the 3rd century AD to impersonate animal spirits and thereby placate them. Various kinds of masks—demon masks, medicine masks, spirit masks—were worn by shamans as they danced to draw into themselves the spirit being addressed, in order to cure an illness or otherwise affect daily life.

Three Kingdoms period. Lack of records makes it impossible to describe accurately dances and dance plays of Korea prior to the period of the Three Kingdoms (*c.* 57 BC–AD 668). Chinese, Japanese, and Korean accounts beginning in the 7th century give some indication of court arts in the Three Kingdoms of Koguryŏ, Paekche, and Silla (Shinla). In Koguryŏ, encompassing what is now Manchuria and northern Korea, Central Asian music and dances were combined with local styles of music and dance. Twelve of 24 pieces in the repertoire were mask dances. So highly regarded were the arts of Koguryŏ that they made up a separate Korean component of the Nine Departments of Musical Art and Dance at the T'ang court in China (25 musical and dance items were identified as Korean), and from the 7th century they were introduced into Japan, where they became the basis of *bugaku* (court masked dance). The strongly Buddhist state of Paekche in the southwest had been in contact with both China and Japan from early in the Christian era. Typical of Paekche was a Buddhist masked-dance processional (*kiak*), originating in southern China and taken to Japan in 612 by a resident of Paekche, Mimaji. No Korean account of *kiak* survives, but Japanese accounts make clear that it was performed as a Buddhist ceremonial for evangelical purposes.

Great Silla period. The third kingdom, Silla, absorbed Koguryŏ and Paekche in the 7th century, and during the Great or Unified Silla period (668–935) the folk and court performing arts of all parts of Korea intermingled. Several major types of masked dance are mentioned in Silla records. The spirit of a noble youth who died to save his father's throne was memorialized in a masked sword dance (before this time, palace dancing girls had performed sword dances, but always unmasked). Masked dances called "The Five Displays" are mentioned in a Silla poetic composition of the 9th century. They included acrobatics, ball juggling, farcical pantomime, shamanistic masked dances, and the lion dance. The similarity of several to Japanese *bugaku* dances has been noted. Others believe "The Five Displays" derive from the "hundred entertainments" of China. Finally, an important dance play honouring Ozoyong, the son of the Dragon God of the Eastern Sea, dates from this period. Ozoyong showed such generosity toward the spirit of plagues that henceforth

The dance play honouring Ozoyong

the spirit promised never to enter a household where a portrait of Ozoyong was hung. Originally derived from animistic beliefs, the dance was modified by Buddhism and was developed in the Chosŏn dynasty (1392–1910) into a spectacular dance play performed by a cast of 5 masked dancers and 16 unmasked dancing girls and accompanied by an ensemble of 37 musicians.

Koryŏ period. The two major court festivals at which performances were held during the Koryŏ period (935–1392) were Buddha's birthday, or the Feast of Lanterns, in the second lunar month, and the midwinter ceremony honouring spirits of local gods. Dances and masked plays from Silla times were carefully preserved and performed on these occasions in a specially decorated and candlelit ceremonial room. New masked plays memorializing loyal warriors who had died in battle were added from the 10th century. Buddha was offered gifts of wine and food, and performance was dedicated to maintaining a reign of peace and harmony. From the time of King Munjong (1046–83), T'ang style dances and sung dramas were performed on other occasions; modified by Korean forms, they became part of Korean court dance in centuries following.

Folk dances and plays undoubtedly go back many centuries before this; in the Koryŏ period, professional troupes also became part of urban life. The practice of court performers holding civil-service jobs in the major cities and in provincial towns probably accounts for the fact that knowledge of court performing arts began to reach beyond the confines of the court during this time. Popular troupes began the process of secularizing religious masked dances (such as the *narye,* which formerly was performed to exorcise evil). They performed acrobatics and shows of skill and at least by the 12th century were staging satiric dialogue plays that held officialdom up to ridicule. (The development of social satire is found in many Asian drama forms: the Vidusaka jester in Sanskrit drama, the god-clown-servants of Indonesian *wayang* shadow plays, and the servants of *kyōgen* comedies in Japan are major roles in these forms.)

Satiric theatre

Chosŏn and modern periods. Buddhism was rejected as a state religion by the Chosŏn (Yi) dynasty (1392–1910), with the result that court entertainments were no longer scheduled according to Buddhist days of worship but at any time court entertainment was required. A Chinese envoy to the Chosŏn court in 1488 described court performances that included the Ozoyong dragon-god dance play, children's dancing, acrobats, ropewalking, and displays of animal puppets. Following invasions by the Japanese (1592) and by the Manchu (1636), court support declined. Former palace performers formed professional troupes, in the process adapting court forms to popular tastes. These performers included all the miscellaneous



Korean female drum dance, sword dance, and masked dance, detail of a watercolour on paper scroll depicting a reception for the governor of P'yŏngyang, Chosŏn dynasty (1392–1910), by Kim Hong-do (18th century). In the National Museum of Korea, Seoul.
By courtesy of the National Museum of Korea, Seoul

stage arts in their repertoire and created from the various court dances and masked plays a type of folk masked play usually termed *sandae togam gug.* A prominent feature was the satiric treatment of depraved Buddhist monks and of grasping officials. Satiric plays were occasionally performed at court as well, but the banishment in 1504 of an actor for ridiculing the institution of kingship in a court play suggests that satire was not welcomed. *P'ansori,* a sung narrative accompanied by virtuoso drumming, was created by professional performers during the Chosŏn period. Either a man or a woman could be the solo singer-dancer, often a shaman. The current repertoire of six long stories was codified in the 19th century by the performer Shin Jae-hyo.

In addition to professional groups, villagers in different areas of the country formed folk groups to perform their own local versions of the *sandae* masked play and dances. Today the *sandae* masked play is performed by villagers in Yangju, Kyŏnggi province, and in South Kyŏngsang province in South Korea and in Pongsan, Hwanghaedo province, North Korea. Performers are males. Masks cover either the whole head or the face and are made from paper or gourds or, occasionally, are carved from wood. They are boldly painted to represent the stock characters of the play: monks, shaman, noblemen, young dancing girl, and others. There may be 20 or 30 masks used; often they are burned and made anew each year to ensure their ritual purity. Performance encompasses singing, dancing, pantomime, and dialogue. The stories enacted vary with the village, but common scenes include offerings to the gods, criticism of venial Buddhist priests, exposure of corruption by gentry and officials, flirtation, and a funeral service that brings absolution. Performances may be given as a rainmaking rite.

Puppet plays

In the Koryŏ period puppet plays were widely performed and very popular among the people. Several types of puppet play developed in Korea. The folk puppet play *Kkoktukaksi,* named after the wife of the main character, is still performed in the summer months in South Korea by farmers in troupes of six or seven players and musicians. Twelve or 15 puppets make a set; they are glove-and-stick figures that can be manipulated by a single puppeteer. One play, with variations, is performed. It consists of eight relatively independent scenes that satirize a figure

Song Kee-yep



The scene of *Aesadang-nori* from Yangju *sandae* mask-dance drama of Korea.

of the gentry who is the major character. Scenes satirizing depraved monks and insulting the gentry, a domestic triangle, and Buddhist prayers for the dead appear to be adapted from masked plays.

Gu gug (literally "old plays") became popular about the middle of the 19th century. They were dramatic songs, danced to gestures and simple group movements. Troupes played throughout the countryside and in the National Theatre, built in Seoul by the government in 1902. Until the 1930s, variety programs of *gu gug* and female court dances were popular entertainments at commercial theatres in the city. Sentimental melodramas, called "new school," or *shimpa,* plays (the same name as in Japan), were performed by a dozen troupes that formed and disbanded between 1908 and about 1930. The new school movement was begun by the novelist Yi Injik. Other major figures had learned the style while studying in Japan. In 1931 the actor Hong Haesŏng and others organized the first drama and cinema exhibition in Korea; later that year its organizers formed the Society for the Study of Dramatic Art, which studied and staged translations of modern European plays. It was active until 1939, when it was suppressed by the Japanese colonial government. Nonetheless, by 1940 about 100 amateur groups were using realistic "new drama" (*singgug*) as a means of social and political protest.

Since World War II. In Korea after 1940 all dramatic groups were obliged to belong to the Japanese-organized Dramatic Association of Korea. Many groups survived the war with Japan by touring small towns and villages. Performances lagged immediately after World War II because of unsettled conditions. A new National Theatre was established in Seoul just before the Korean War began; national support included subsidies for performances. In both North and South Korea virtually all theatres were destroyed by the war. Excellent theatres were constructed in the 1970s and '80s, however, and performances are numerous in both political areas.

In South Korea the National Theatre supports large-scale musical dramas, folk dance, and traditional music through performance and troupe subsidies. Among semiprofessional little theatre groups the Drama Center, Jayu (Free), Minye (Folk), Silhom (Experimental), and Kagyŏ (Bridge) theatre troupes are well established. Social problems and the integration of traditional and modern ways are common themes in contemporary plays. Western-style opera, ballet, and modern-dance troupes also perform.

Plays in North Korea are required to represent socialist construction, be nationalistic, and offer the masses pleasure, following the precepts of "self-reliance" (*juche*) of President Kim Il-sung (1912–94). A small number of "model" works emphasizing music or dance within grandiose spectacles ("Song of Glory" has a cast of 5,000) make up the repertoire of major theatres.

JAPAN

Among the most varied and technically complex theatre arts in Asia are those of Japan. Music and dance gradually evolved into highly developed dramatic and theatrical forms, the most important of which are nō dance drama, popular kabuki theatre, and bunraku puppet drama.

Formative period. From prehistoric times, dances have served as an intermediary between man and the gods in Japan. *Kagura* dances dedicated to native deities and performed at the Imperial court or in villages before local Shintō shrines are in essence a symbolic reenactment of the propitiatory dance that lured the sun goddess Amaterasu from the cave in ancient myth. Although *kagura* dance has been influenced by later more sophisticated dance forms, it is still performed much as it was 1,500 years ago, to religious chants accompanied by drums, brass gongs, and flutes. At the same time, villagers had their rice-planting dances, performed either at New Year's as a prayer for good planting or during the planting season in early summer. These lively dances were later, in the 14th century, brought to the cities and performed as court entertainment and called *dengaku* ("field music").

7th to 16th centuries. A massive influx into Japan of Chinese and Korean arts and culture occurred between the 6th and the 10th centuries. It has been noted that a Korean performer, Mimaji (Mimashi in Japanese), brought the Buddhist *gigaku* processional dance play to the Japanese court in 612. Mimashi established an official school to train Japanese dancers and musicians in *gigaku.* Other Korean and Chinese performers from Paekche and Koguryŏ were invited in following years. *Gigaku* masks cover the entire head (as do Korean folk masks today). Carved of wood and painted with lacquer, the 223 masks that remain (most in the Shōsō Temple in Nara) date back as early as the 7th century. They are superb examples of the art of mask making, strong-featured and beautifully conceived. From a description of a 13th-century performance, *gigaku* apparently consisted of a succession of scenes enacted as characters passed by. Masks characterized an Aryan-featured dignitary called Baramon (or Brahman, indicating Indian origin), a fierce wrestler, a Buddhist monk, a princess of the state of Wu in China, a bully, a wistful old man, and others. Some scenes were serious, others were earthy slapstick.

The *gigaku* processional dance play

Bugaku court dances introduced from Korea also were patronized by the court. They supplanted *gigaku* as official court entertainment, and *gigaku* disappeared as a performing art by the 12th century. It was the custom to have performers of *bugaku* enter from dressing rooms to the right and the left of the raised platform stage: "right" dances, costumed in orange or red, were those from India, Central Asia, or China proper; "left" dances, costumed in blue-green, were those from Korea and Manchuria. *Bugaku* is usually performed by groups of four, six, or eight male dancers who move in deliberate, stately steps, repeating movements in the four cardinal directions. Musical accompaniment is by drums, bells, flute, lute, and *shō* (panpipe). A composition consists of three sections: introduction, development, or "scattering," and speeding up (*jo-ha-kyū*). Japanese performers and courtiers created new compositions within the old style in the 10th and 11th centuries. Still, *bugaku* represents a remarkable preservation of ancient Chinese, Indian, and Korean music and dance that have long since disappeared in their countries of origin. *Bugaku* has been performed by musicians attached to the Imperial court and to major Shintō shrines from the 7th century without break to the present day.

Performance of *bugaku* court dances

Juggling, acrobatics, ropedancing, buffoonery, and puppetry—the "hundred entertainments" of China and called *sangaku,* "variety arts," in Japan—became widely popular as well. During the Heian period (794–1185) professional troupes, ostensibly attached to temples and shrines to draw crowds for festival days, combined these lively stage arts, now called *sarugaku* (literally, monkey or mimic music), with dancing to drums from *dengaku* and began to perform short plays consisting of alternate sections of dialogue, mimicry, singing, and dancing. Sometime in the 14th century a *sarugaku* actor from Nara named Kan'ami incorporated in his plays a chanted dance (*kuse-mai* or *kōwaka-mai*), for the first time creating the possibility of dramatic dance that could carry forward a story. This

ZEFA—G. Haasch



Bugaku, a court dance adapted to Japanese tastes from the dance and music of 8th-century China and Korea.

fusion of dance, drama, and song, which soon came to be known as *sarugaku-no-nō*, or simply nō, marked a revolutionary advance in Japanese theatrical art. Kan'ami's son, Zeami, refined the style of performance, composed 50 or more of the finest nō plays in the repertoire, and wrote fundamental treatises on the art of acting and dramaturgy.

When Zeami was 11, the military ruler of Japan, the shogun Ashikaga Yoshimitsu, saw him perform, became enamoured of the boy's beauty, and took him into his residence in Kyōto as a companion. For most of his life, Zeami benefited from the patronage and the refined audiences that stemmed from this circumstance.

Nō theatre

The borrowings of nō from other arts are many. The exquisite masks for which nō theatre is famous have a quality of serenity, a neutrality of expression that places them in a rank perhaps unmatched in the world. Yet, historically there is no doubt that they are derived from earlier *bugaku* and *gigaku* masks and hence are related, if distantly, to the masks of Korea, China, and India. One evidence of the special development of nō masks is that they are smaller than previous masks; they cover only the face proper. From *bugaku* music, Zeami took the three-part structure of the nō drama. A normal nō program consists of five plays, which are grouped into three dramatic units: the introduction, the development, and the conclusion. The first play, a "god" play, constitutes the introduction; the second drama, or "warrior" play, is the introduction of the development section; the third, or "woman," play is the development of the development; the fourth, or "living person," play is the conclusion of the development; and the fifth, or "demon," play is the conclusion. Drums and flute were taken over from earlier musical forms, and nō chanting grew out of Buddhist prayer chants. The songs' poetic meter of alternating phrases of seven and five syllables had come from China six centuries earlier and was the standard Japanese poetic form. On the other hand, the nō stage represents an advance on the simple square platform of *bugaku*. A sharply peaked roof over the stage is supported by four pillars—to help the performer orient himself as he looks out through tiny eye holes in the mask—and a long ramp, *hashigakari*, emphasizes the entrance of major characters.

Typical of a number of nō plays that are dramatizations of Chinese history and legends is the 15th-century *Yokihi*, by Komparu Zenchiku, based on the 9th-century narrative poem *Ch'ang hen ko* ("The Song of Everlasting Sorrow") by Po Chü-i. The original describes Emperor Hsüan-tsung's love for his concubine Yang Kuei-fei (Yōkihi in Japanese). The nō play emphasizes the Buddhist sentiment of the evanescence of mortal life and the inevitability of pain and sadness. Every nō play contains Chinese poems, quoted verbatim or paraphrased so as to appeal to the educated spectator. It was a first principle of dramatic writing, said Zeami, to base a play on a well-known incident in which the central character was familiar to the audience. Zeami's plays emphasized the quality of restrained beauty (*yūgen*), a concept derived in part from Zen Buddhism. Later plays, especially those by Kanze Kojirō Nobumitsu (1435–1516), such as *Momijigari* (*The Maple Viewing*) and *Ataka* (*The Ataka Barrier*), emphasize action and spectacle (*fūryū*).

The *kyōgen* farce comedy

On the usual nō program, each play was followed by a *kyōgen* farce comedy. The antecedents of *kyōgen* cannot be described with certainty, but it is probable that *kyōgen*'s short sketches of master-servant quarrels, husband-wife arguments, animal fables, and scenes of rustic life derive from early *sangaku* entertainments. A few *kyōgen* plays are accompanied by the drums and flute of nō. The ritual play *Okina*, performed as an auspicious prayer for longevity at the beginning of a nō-*kyōgen* program, is in both repertoires, and some suggest that the *kyōgen* version is the older. The style of *kyōgen* music (*koutai*) is distinct from that of nō music; it is derived directly from popular songs. *Kyōgen* plays with music are, however, a rarity. The usual play is a straight dialogue drama, making it perhaps the oldest developed form of nonmusical play in East Asia. Dialogue is composed in colloquial language of the 15th century, in short phrases suitable for comedy. Movement is highly stylized, again for comic effect. Masks may be worn for the roles of animals and demons, but most roles are played unmasked. *Kyōgen* texts do not seem to have been committed to writing until the 16th century, suggesting that actors traditionally ad-libbed their parts. Today *kyōgen* actors commit lines to memory.

Azuchi-Momoyama period. Nō and *kyōgen* were dance and theatre forms that had come to express the gravity and decorum of a rigidly formal samurai ruling class by the end of the Azuchi-Momoyama period (1574–1600). Artistically severe and highly disciplined, nō was imbued with the sternly pessimistic philosophy of Buddhism. In content, nō plays taught the folly of worldly power and position, that time destroys all living things. The heroes of play after play pray for the divine intercession of Amida (Amitābha) in order that they, tormented ghosts of dead warriors and court ladies, may break free of earthly attachments and achieve Buddhist salvation. In contrast to this,

Photograph by Kunihei Kameda



Scene from Yuya, a woman nō play attributed to Zeami, showing (left foreground) the *shite* (principal actor) and (right) *waki* (supporting actors). The *hayashi* (musicians) are seated in front of the pine tree painted on the *kagami-ita* (rear wall); the *jultai* (chorus) sits at right; and a *kokeu* (stage assistant) sits at the left. A portion of the *hashigakari* (ramp leading to stage) is at far left.

commoners of the cities in the late 16th century began to perform their own dances and plays that were up-to-date, lively, exciting, and at times morally licentious. They were intended to appeal to literate townsmen, well-to-do wives of merchants, workers, and the fops, wits, and dandies of the burgeoning cities.

Tokugawa period. During the Tokugawa period (1603–1867) nō was assiduously cultivated by samurai as a refined accomplishment. Commoners were forbidden by law to study nō and were excluded from performances except on special "subscription" occasions, when any person, high or low in rank, could see nō performed outdoors in a large enclosure. Nō became the exclusive theatre art of the warrior class, while *bugaku* continued as the chief performing art of the Imperial court.

The earliest important urban entertainments of the commoner in Japan were secularized forms of Buddhist dance plays (*ennen*) and folk dances (*yayako odori* and *kaka odori*) that came to be called *fūryū* ("drifting on the wind") dances. They were enormously popular.

Origin of kabuki

In 1603 several kinds of urban dances were arranged by a young woman named Okuni into a new dance, called kabuki. Other troupes of female prostitute-performers adopted the sensuous and popular style of Okuni's kabuki dance. A scroll of the period shows Okuni as a young, fashionably dressed samurai, indolently leaning on a sword, dallying with a teahouse girl. Around her neck hangs a Christian crucifix, not as a religious article but as an exotic decoration recently introduced by "southern barbarian" Portuguese merchants. Other pictures of the time show young women playing the three-stringed samisen as they recline sensuously on tiger skins, dancing girls circling about them. Audiences of monks, warriors, young lovers, and townsmen gaze raptly at this appealing and even bizarre sight. The original sensuous appeal of women's kabuki continued long after women were banned from the stage in 1629. From about Okuni's time until 1652, troupes of boy prostitutes performed graceful kabuki dances to samisen music to attract customers. The appearance of professional women and boy performers in kabuki was a phenomenon of urban society.

In 1653, when the authorities required kabuki to be performed by adult males, kabuki began to develop as a serious art. During the Genroku era (1688–1703), most of kabuki's essential characteristics were established. Large, commercial theatre buildings holding several thousand spectators were constructed in the three major cities—Edo (Tokyo), Kyōto, and Ōsaka. The stage, which previously had been simply a copy of the nō stage, became wider and deeper and was equipped with a draw curtain to separate acts; and in the early 1700s a ramp (*hanamichi*) was constructed from the rear of the auditorium to the stage for actors' entrances and exits. The idea of the rampway came from the nō *hashigakari*, but, in typical kabuki fashion, it was transformed into an infinitely more theatrical device. From the puppet theatre, kabuki borrowed the use of fairly elaborate scenery, the revolving stage (100 years before its use in Europe), traps, and lifts. To the old nō drums and flute were added the new samisen, a large drum, a dozen bells, cymbals, gongs, and two types of wooden clappers, making the resulting music flexible and varied.

Nō and *kyōgen* plays were often performed as kabuki in the early decades. A print from about 1670 shows kabuki actors performing *Sumidagawa* (*The Sumida River*), with costumes and properties modeled closely on the nō original. But it was not considered proper for "beggars of the riverbed" to stage the art of the warrior class. By Genroku times, new kabuki dramatic styles had emerged. The actor Sakata Tōjūrō (1647–1709) developed a relatively realistic, gentle style of acting (*wagoto*) for erotic love stories in Kyōto, while in Edo, a stylized, bravura style of acting (*aragoto*) was created by the actor Ichikawa Danjūrō I (1660–1704) for bombastic fighting plays. In the play *Sukeroku yukari no Edo zakura* (*Sukeroku: Flower of Edo*) written by Tsuuchi Jihei II in 1713, the two styles are blended most successfully. The hero, Sukeroku, is a swaggering young dandy and lover acted largely in the Edo style, while Sukeroku's brother, Shimbei, is a meek, gently comic foil in the Kyōto style. Genroku-period kabuki plays are lusty and active and contain much verbal and physical humour.



A classic *mie* in the *aragoto* style. Ichikawa Danjūrō XII in the role of the warrior hero in *Shibaraku*, a kabuki play written by Ichikawa Danjuro I and first performed in 1697.

By courtesy of SHOCHIKU CO., LTD., photograph by Fumio Watanabe

Kabuki theatres were required to be built in special entertainment quarters (near licensed quarters for prostitution), along with puppet theatres. Puppets, imported from Korea centuries earlier, were fused with epic storytelling and the resulting narrated play accompanied by samisen music sometime before 1600. The earliest tales were about the princess Jōruri (hence, *jōruri* as another name for puppet plays). This and other legends were in the nature of Buddhist miracle stories, the obligatory scene being one in which Buddha sacrifices himself or otherwise brings to life one of the main characters. Simple doll puppets, held overhead by one man, animated these blood-and-thunder *kojōruri* ("old *jōruri*") puppet plays.

The role of Chikamatsu in the puppet theatre

A new style of puppet play was created in 1686 by the writer Chikamatsu Monzaemon (1653–1725) and the chanter Takemoto Gidayū at the Takemoto Puppet Theatre in Ōsaka, the city which became the home of puppet theatre in Japan. The chanter is responsible not only for narrating the play but for providing the voices of all the puppet characters as well; Gidayū's expressive delivery style remains influential to this day. Chikamatsu went on to become Japan's most famous playwright. Although he is best known for his puppet plays, he wrote kabuki plays as well, most of them for Sakata Tōjūrō. From Tōjūrō, Chikamatsu learned the soft style of kabuki performance and the situation that is so unique to early kabuki, in which a comic lover visits a courtesan in the licensed quarter and quarrels with her. Between *Sonezaki shinjū* (*Love Suicides at Sonezaki*), written in 1703, and *Shinjū ten no Amijima* (*Love Suicides at Amijima*), written in 1721 three years before his death, Chikamatsu composed for the puppet theatre a dozen domestic tragedies handling the theme of lovers' suicide. As early as 1678, kabuki plays were dramatizing current city scandals, lovers' suicides, murders, and tragic deaths. One of the most characteristic features of kabuki was its contemporaneous dramatic subject matter; puppet drama was much changed when Chikamatsu brought this quality from kabuki into his puppet plays.

The puppet theatre underwent significant physical change when the puppet operators, samisen player, and chanter were made fully visible to the audience in 1705. In the 1720s and '30s puppet plays gradually became more dramatic and less narrative under the influence of kabuki. A revolutionary three-man puppet was created in which

mouth, eyes, eyebrows, and fingers could move, encouraging writers to compose dramatic plays calling for complex emotional expression. A theatre manager and writer, Takeda Izumo II (1691–1756), collaborated with several other authors on all-day history plays, the so-called "Three Great Masterpieces" of puppet drama: *Sugawara denju tenarai kagami* (1746; *The Secret of Sugawara's Calligraphy*), *Yoshitsune sembonzakura* (1747; *Yoshitsune and the Thousand Cherry Trees*), and *Kanadehon chūshingura* (1748; *Chūshingura: The Treasury of Loyal Retainers*). The latter is the best-loved and most often performed drama ever written in Japan; it typifies mature puppet drama. It is based on actual events that occurred from 1701 to 1703: 46 retainers avenged the death of their lord by killing his enemy and were then sentenced to commit *seppuku*, or ritual suicide by disembowelment. (A 47th conspirator was not involved in the actual killing.) The major scenes of the suicides of the lord Hangan and his retainer Kampei are intensely emotional scenes of self-sacrifice. Such scenes normally occur as the final section of the third act in a five-act history play and are called *sewa* ("family") scenes because, although the figures are samurai, tearful family separation is the emphasis of the scene. *Ichinotani futaba gunki* (1751; *Chronicle of the Battle of Ichinotani*) contains a *migawari* ("child substitution") scene, typical of puppet history plays, which is, if anything, even more tear-provoking: in response to the wishes of his lord Yoshitsune, General Kumagai slays his own son, so that the son's head may be substituted for that of a prince who has been condemned to die. Although the puppet plays' emotionalism and lack of humour were foreign to kabuki, the serious dramas were so popular with audiences that they were adopted, with changes, to kabuki. Today the best puppet plays are equally a part of the kabuki and puppet theatre repertoires.

19th-century kabuki plays

During the 19th century the most important kabuki dramas were written in Edo, by Tsuruya Namboku IV and Kawatake Mokuami. They wrote all the standard types of kabuki play—*sewamono* (domestic), *jidaimono* (history), and *shosagoto* (dance plays)—in large numbers; each wrote between 150 and 200 plays in his professional career. They spent their lives in the kabuki theatre as writers. Although neither was formally educated, their plays reflect with great discernment the desperate social conditions that prevailed as the feudal system in Japan neared its collapse. Thieves, whores, murderers, pimps, and ruthless masterless samurai are major figures in a new type of play, *kizewamono*, or gangster play, which Namboku created and Mokuami developed. They wrote for the talents of star actors: Namboku wrote for the finest *onnagata* (female impersonator) of his time, Iwai Hanshirō V, and Mokuami wrote for Ichikawa Danjūrō IX and a remarkable actor of gangster roles, Ichikawa Kodanji IV. Each was a master of kabuki art, and between them they added new dimensions to kabuki's stylized form. Namboku created rhythmic dialogue composed in phrases of seven and five syllables; Mokuami used puppet-style music to heighten the pathos of certain scenes and wrote elaborately conceived major speeches which required exceptional elocutionary skill on the part of the actor.

Meiji period. Nō, puppet theatre, and kabuki were affected in differing degrees by the abolition of feudalism in 1867. At a stroke, the samurai class was eliminated and nō lost its base of economic support. Important actors retired to the country to eke out a living as menial workers. For several years nō was not performed at all, except that Umewaka Minoru, a minor actor, gave public performances in his home and elsewhere between 1868 and 1876. In 1881 a public stage was built in Shiba Park, Tokyo, for performances sponsored by the newly formed Nō Society and by its successor, the Nō Association. The most influential supporter of nō during the Meiji period (1868–1912) was the aristocrat Iwakura Tomomi. The study of nō came to be a highly regarded activity among the middle classes, and in time each of the five nō schools (Kanze, Hōshō, Komparu, Kongō, and Kita) became financially stable, sponsoring their own performances and building their own theatres in the major cities.

The end of feudal society forced nō to seek and cultivate a new audience; the popular audience of kabuki and the puppet theatre, however, continued with little change during the Meiji period. Kabuki audiences remained large and loyal, but audiences for puppet plays continued to decline as they had for the previous hundred years. There was a brief revival of interest in Ōsaka puppet drama in the 1870s under the impetus of the theatre manager Daizō, the fourth Bunrakuken, who called his theatre Bunrakuza (from the name of a troupe organized by Uemura Bunrakuken early in the century). The popular term for puppet drama, bunraku, dates from this time. Learning to chant puppet texts became a vogue during the late Meiji period. In 1909 the Shōchiku theatrical combine supported performances at the Bunraku Puppet Theatre in Ōsaka, and by 1914 this was the only commercial puppet house remaining.

As they always had, kabuki writers and actors of the Meiji period tried to place current events on the stage. Thus, the actor Onoe Kikugorō V began acting in a series of contemporary plays, dressed in daily kimono or Western clothes and with his hair cut Western fashion (the origin of *zangirimono*, or the so-called "cropped-hair plays"), in the late 19th century. Western influence also was seen in theatre construction, with the first European-style theatre built for kabuki in Tokyo in 1878. Released from previous government restrictions, kabuki artists created dance dramas from the nō play *The Maple Viewing* and others, in which the elevated tone of the nō original was purposely retained. Kabuki attendance was more than a million

Hiroshi Kaneko



Bunraku, a scene from the comedy *Tsuri onna* ("Fishing for a Wife"), a puppet performance of a kabuki dance version of the *kyōgen* original play. Chief puppeteers wear conventional dress, minor puppeteers wear black, hooded costumes. Pine-and-plank scenery indicates the nō-*kyōgen* origin of the play.

spectators yearly. But, in spite of prosperity and seeming adaptation to new conditions, by the early decades of the 20th century, new artistic creation in kabuki reached an impasse, and thereafter kabuki became restricted almost as much as bunraku and nō to a classic repertoire of plays.

Scholars and artists, learning of Western drama, organized successive groups designed to reform kabuki—that is, to eliminate excessive stylization and to press for a more realistic manner of performance. The actor Ichikawa Danjūrō IX acted in historically accurate (and reportedly dull) *katsureki geki* ("living history" plays) written by the journalist Fukuchi Ōchi. Three *shin* kabuki ("new kabuki" plays) written by the scholar Tsubouchi Shōyō were influenced by Shakespeare, whose plays Tsubouchi was then translating. In 1908 a young actor, Ichikawa Sadanji II, returned from a year's study and observation in Europe. These and other influences produced few long-lasting changes in kabuki, but they did set the stage for the creation of new kinds of drama that would depart radically from traditional forms.

Contributions of Kawakami

The first plays in Japan consciously based on Western models were those arranged and acted in by Kawakami Otojirō. Kawakami's first plays were political and nationalistic in intent. After he and his wife Sada Yakko had performed in Europe and America (1899 and 1902), they introduced to Japan adaptations of Shakespeare, Maurice Maeterlinck, and Victorien Sardou. These *shimpa,* or "new school," plays, however, were little more than crude melodramas. Yakko and other actresses performing in *shimpa* marked the first time women had appeared on the professional stage since Okuni's time. One *shimpa* troupe continues to perform today, in a style that retains turn-of-the-century sentiment and mannerisms.

In 1906 the Literary Society was established by Tsubouchi Shōyō to train young actors in Western realistic acting, thus beginning the serious study of Western drama. The first modern play (*shingeki*) to be staged in Japan in the Western realistic manner was Ibsen's *John Gabriel Borkman,* directed by Osanai Kaoru in 1909 at his Free Theatre, which was modeled on the "free theatres" of Europe. Much to the detriment of *shingeki*'s development, major European playwrights—George Bernard Shaw, Anton Chekhov, Leo Tolstoy, Gerhart Hauptmann, Maeterlinck—were chosen for production over aspiring Japanese authors by all the important early troupes: the Art Theatre (1913–19) founded by a Tsubouchi disciple, Shimamura Hōgetsu; the Stage Association; and the Tsukiji Little Theatre (1924–28). The members of *shingeki* troupes were earnest amateurs, strongly motivated by artistic and social ideals to create a theatre that reflected life in 20th-century Japan. The only early *shingeki* troupe to survive World War II was the Literary Theatre (1937).

Since World War II. Following Japan's surrender in 1945, kabuki and bunraku plays that the American occupation forces considered feudal, such as *Kanjinchō* (*The Subscription List*) and *Chūshingura: The Treasury of Loyal Retainers,* were banned briefly. Since then, nō and kabuki have greatly prospered, while bunraku has become increasingly subsidized. Modern drama authors and performers, many of whom had been jailed or persecuted by Japanese authorities during World War II for liberal and leftist beliefs, were encouraged by the occupation forces. Important *shingeki* troupes founded in the immediate postwar years include the Actors' Theatre (1944), directed by Senda Koreya, an expert on the works of Bertolt Brecht; The People's Theatre, devoted to progressive social and political issues; and Theatre Four Seasons (1953), which specialized first in French drama and later in American musicals. The full range of Japanese modern life was examined in such *shingeki* plays as Kinoshita Junji's nostalgic folk drama *Yūzuru* (1949; *Twilight Crane*); Mishima Yukio's psychological study of cruelty *Sado koshaku fujin* (1965; *Madame de Sade*); Tanaka Chikao's *Maria no kubi* (1959; *The Head of Mary*), about the atomic bombing of Hiroshima; the Social Realist play *Kazanbaichi* (1938; *Land of Volcanic Ash*) by Kubo Sakae; and Inoue Hisashi's comic tribute to popular theatre, *Keshō* (1983; "Makeup").

Shingeki's orthodox realism, its increasing commercialism, and its impotence during the struggle to block the 1960 passage of the United States–Japan Treaty of Mutual Cooperation and Security alienated younger theatre artists. In the 1960s, for both political and artistic reasons, director-authors Suzuki Tadashi, Terayama Shūji, Kara Jūrō, and Ohta Shōgo formed their own acting companies in order to create unique new theatrical works incorporating stylized acting, song, dance, and brilliant stage effects. They believed that it was necessary to turn back to traditional Japanese culture and arts in order to move forward toward a Japanese theatre unfettered by Western models. Social disjuncture and alienation are common themes of the absurdist plays *Tomodachi* (1967; *Friends*), by Abe Kōbō, Betsuyaku Minoru's *Zo* (1962; *The Elephant*), and Satoh Makoto's *Atashi no Beatles* (1967; *My Beatles*).

The most extreme rejection of both Western mimesis and traditional Japanese aesthetics is seen in *butō* (or *ankoku butō,* "dance of darkness"), a postmodern movement begun by Hijikata Tatsumi and Ohno Kazuo in the 1960s in which formal dance technique is eschewed and primal sexuality and the grotesque are explored. The *butō* troupes Sankaijuku, Dairakudakan, and Byakko-sha, as well as individual dancers such as Tanaka Min, often tour Europe and the United States.

In some ways the effects of modernization on the performing arts in 20th-century Japan have been great. In the 1950s the country's motion-picture industry was the second largest in the world, only to be displaced by television, which saturated every corner of Japan by the end of the 1960s. Yet attendance for live theatre did not decline; on the contrary, in the general affluence of Japanese society of the 1970s and '80s, attendance continued to grow at almost every type of performance. Overall in Tokyo, some 3,000 live theatre productions are staged annually, and a boom in theatre building has added scores of elegant new playing spaces for both traditional and avant-garde performance. Nō and kabuki dance are avidly studied by thousands of amateurs; three new national theatres (built 1966–85) house subsidized productions of nō, kabuki, and bunraku; lavish theatre and dance festivals annually host local and foreign troupes; and international tours regularly introduce Japanese plays and dances to foreign audiences. Live theatre of all types flourishes in Japan, each form appealing to its own sector of the overall audience.

(J.R.B.)

BIBLIOGRAPHY

Visual arts. *China:* General works include LAURENCE SICKMAN and ALEXANDER SOPER, *The Art and Architecture of China,* 3rd ed. (1968, reissued 1991); DIETRICH SECKEL, *The Art of Buddhism,* rev. ed. (1968; originally published in German, 1962); OSVALD SIRÉN, *A History of Early Chinese Art,* 4 vol. (1929–30, reprinted 4 vol. in 2, 1970); MICHAEL SULLIVAN, *The Arts of China,* 3rd ed. (1984); and CRAIG CLUNAS, *Superfluous Things: Material Culture and Social Status in Early Modern China* (1991).

Archaeology and Neolithic and Bronze Age arts are considered in ROBERT W. BAGLEY, *Shang Ritual Bronzes in the Arthur M. Sackler Collections* (1987); KWANG-CHIH CHANG, *The Archaeology of Ancient China,* 4th ed., rev. and enlarged (1986); WEN FONG (ed.), *The Great Bronze Age of China* (1980); BERNHARD KARLGREN, *A Catalogue of the Chinese Bronzes in the Alfred F. Pillsbury Collection* (1952); THOMAS LAWTON, *Chinese Art of the Warring States Period* (1983); LI CHI (CHI LI), *Anyang* (1977); MAX LOEHR, *Ritual Vessels of Bronze Age China* (1968); J.A. POPE *et al., The Freer Chinese Bronzes,* 2 vol. (1967); JESSICA RAWSON, *Ancient China: Art and Archaeology* (1980), and *Western Zhou Ritual Bronzes from the Arthur M. Sackler Collections,* 2 vol. (1990); and CHARLES D. WEBER, *Chinese Pictorial Bronze Vessels of the Late Chou Period* (1968).

Descriptions of architecture and gardens are found in ANDREW BOYD, *Chinese Architecture and Town Planning, 1500 B.C.–A.D. 1911* (1962); JOHN HAY, *Kernels of Energy, Bones of Earth: The Rock in Chinese Art* (1985); JI CHENG (CH'ENG CHI), *The Craft of Gardens,* trans. from Chinese (1988); MAGGIE KESWICK, *The Chinese Garden* (1978, reissued 1986); RONALD G. KNAPP, *The Chinese House: Craft, Symbol, and the Folk Tradition* (1990); LIANG SSU-CH'ENG (SSU-CH'ENG LIANG), *A Pictorial History of Chinese Architecture,* ed. by WILMA FAIRBANK (1984); JOHANNES PRIP-MOLLER, *Chinese Buddhist Monasteries,* 2nd ed. (1967); OSVALD SIRÉN, *The Walls and Gates of Peking* (1924), *The Imperial Palaces of Peking,* 3 vol. (1926, reprinted 1976), and *Gardens of China* (1949); ROLF A. STEIN, *The World*

in Miniature: Container Gardens and Dwellings in Far Eastern Religious Thought (1990; originally published in French, 1987); NANCY SHATZMAN STEINHARDT, *Chinese Imperial City Planning* (1990), and *Chinese Traditional Architecture* (1984); and PAUL WHEATLEY, *The Pivot of the Four Quarters: A Preliminary Enquiry into the Origins and Character of the Ancient Chinese City* (1971).

Ceramics are dealt with in CÉCILE BEURDELEY and MICHEL BEURDELEY, *A Connoisseur's Guide to Chinese Ceramics* (1974; originally published in French, 1974); MARGARET MEDLEY, *The Chinese Potter: A Practical History of Chinese Ceramics*, 3rd ed. (1989); SUZANNE KOTZ (ed.), *Imperial Taste: Chinese Ceramics from the Percival David Foundation* (1989); W.B.R. NEAVEHILL, *Chinese Ceramics* (1975); MARY TREGEAR, *Song Ceramics* (1982); SUZANNE G. VALENSTEIN, *A Handbook of Chinese Ceramics*, rev. and enlarged ed. (1989); and WILLIAM WATSON, *Tang and Liao Ceramics* (1984).

Among the numerous works on painting and calligraphy, the following may be recommended: WILLIAM REYNOLDS BEAL ACKER (ed. and trans.), *Some T'ang and Pre-T'ang Texts on Chinese Painting*, 2 vol. in 3 (1954–74), and a reprint of vol. 1 (1979); RICHARD M. BARNHART, *Peach Blossom Spring: Gardens and Flowers in Chinese Painting* (1983); SUSAN BUSH, *The Chinese Literati on Painting: Su Shi (1037–1101) to Tung Ch'i-ch'ang (1555–1636)* (1971); SUSAN BUSH and CHRISTIAN MURCK (eds.), *Theories of the Arts in China* (1983); SUSAN BUSH and HSIO-YEN SHIH (compilers and eds.), *Early Chinese Texts on Painting* (1985); JAMES CAHILL, *Chinese Painting* (1960, reissued 1985), *Hills Beyond a River: Chinese Painting of the Yüan Dynasty, 1279–1368* (1976), *Parting at the Shore: Chinese Painting of the Early and Middle Ming Dynasty, 1368–1580* (1978), *The Distant Mountains: Chinese Painting of the Late Ming Dynasty, 1570–1644* (1982), and *The Compelling Image: Nature and Style in Seventeenth-Century Chinese Painting* (1982); VICTORIA CONTAG and WANG CHI-CH'IEN, *Seals of Chinese Painters and Collectors of the Ming and Ch'ing Periods*, rev. ed. (1966); TSENG YU-HO ECKE (YU-HO TSENG), *Chinese Calligraphy* (1971); WEN C. FONG, *Beyond Representation: Chinese Painting and Calligraphy, 8th–14th Century* (1992); MARILYN FU and SHEN FU, *Studies in Connoisseurship*, 3rd ed. (1973), paintings from the Ming and Ch'ing dynasties; SHEN FU, *Traces of the Brush: Studies in Chinese Calligraphy* (1977); ROGER GOEPPER, *The Essence of Chinese Painting* (1963); R.H. VAN GULIK, *Chinese Pictorial Art as Viewed by the Connoisseur* (1958, reprinted 1981); WAI-KAM HO *et al., Eight Dynasties of Chinese Painting* (1980); LOTHAR LEDDEROSE, *Mi Fu and the Classical Tradition of Chinese Calligraphy* (1979); SHERMAN E. LEE and WAI-KAM HO, *Chinese Art Under the Mongols: The Yüan Dynasty, 1279–1368* (1968); CHU-TSING LI, *The Autumn Colors on the Ch'iao and Hua Mountains: A Landscape by Chao Meng-fu* (1965); CHU-TSING LI (ed.), *Artists and Patrons: Some Social and Economic Aspects of Chinese Painting* (1989); MAX LOEHR, *The Great Painters of China* (1980); KIYOHIKO MUNAKATA (ed. and trans.), *Sacred Mountains in Chinese Art* (1991); ALFREDA MURCK and WEN C. FONG (eds.), *Words and Images: Chinese Poetry, Calligraphy, and Painting* (1991); CHRISTIAN F. MURCK (ed.), *Artists and Traditions: Uses of the Past in Chinese Culture* (1976); YŪJIRŌ NAKATA (ed.), *Chinese Calligraphy*, trans. from Japanese and adapted by JEFFREY HUNTER (1983); *Shodō Zenshū*, 28 vol. (1954–68), a collection of calligraphy; JEROME SILBERGELD, "Chinese Concepts of Old Age and Their Role in Chinese Painting, Painting Theory, and Criticism," *Art Journal*, 46(2):103–114 (Summer 1987), "Chinese Painting Studies in the West: A State-of-the-Field Article," *Journal of Asian Studies*, 46(4):849–897 (1987), and *Chinese Painting Style* (1982); OSVALD SIRÉN, *Chinese Painting: Leading Masters and Principles*, 7 vol. (1956–58, reissued 1974); MICHAEL SULLIVAN, *The Birth of Landscape Painting in China*, 2 vol. (1962–80), and *Chinese Landscape Painting: The Sui and T'ang Dynasties* (1980); FRITZ VAN BRIESSEN, *The Way of the Brush: Painting Techniques of China and Japan* (1962, reissued 1978); MARSHA WEIDNER (ed.), *Flowering in the Shadows: Women in the History of Chinese and Japanese Painting* (1990); and YU FEIAN (FEI-AN YÜ), *Chinese Painting Colors: Studies on Their Preparation and Application in Traditional and Modern Times*, trans. from Chinese by JAMES SILBERGELD and AMY MCNAIR (1988).

Decorative arts are presented in NANCY ZENG BERLINER, *Chinese Folk Art: The Small Skills of Carving Insects* (1986); JESSICA RAWSON and JOHN AYERS, *Chinese Jade Throughout the Ages* (1975); OSVALD SIRÉN, *Chinese Sculpture from the Fifth to the Fourteenth Century*, 4 vol. (1925, reprinted 1970); ZHOU ZUN (HSÜN CHOU) and GAO CHUNMING (CH'UN-MING KAO), *5000 Years of Chinese Costumes* (1987; originally published in Chinese, 1984); and WANG SHIXIANG (SHIH-HSIANG WANG), *Classic Chinese Furniture: Ming and Early Qing Dynasties* (1986; originally published in Chinese, 1985).

Analyses of 20th-century arts include JOAN LEBOLD COHEN, *The New Chinese Painting, 1949–1986* (1987); ROBERT HATFIELD ELLSWORTH, *Later Chinese Painting and Calligraphy, 1800–1950*, 3 vol. (1987); ELLEN JOHNSTON LAING, *The Winking Owl: Art in the People's Republic of China* (1988); CHU-TSING LI, *Trends in Modern Chinese Painting* (1979); JEROME SILBERGELD and GONG JISUI (JISUI GONG), *Contradictions: Artistic Life, the Socialist State, and the Chinese Painter Li Huasheng* (1993); and MICHAEL SULLIVAN, *Chinese Art in the Twentieth Century* (1959), and *The Meeting of Eastern and Western Art*, 2nd ed. (1989). (Je.Si.)

Korea: Survey studies include ANDREAS ECKARDT, *A History of Korean Art* (1929; originally published in German, 1929); CHEWŎN KIM and WON-YONG KIM (eds.), *Korean Arts*, 3 vol. (1956–63); CHEWŎN KIM and WON-YONG KIM, *Treasures of Korean Art* (1966); EVELYN MCCUNE, *The Arts of Korea* (1962); *The Arts of Korea*, 6 vol. (1979), on prehistoric art, painting, Buddhist art, ceramics, handicrafts, and architecture; CHEWŎN KIM and LENA KIM LEE (I-NA KIM), *Arts of Korea* (1974); YI KI-BAEK (KI-BAEK YI), *5,000 Years of Korean Arts* (1976); KIM WON-YONG (WON-YONG KIM), *Art and Archaeology of Ancient Korea* (1986); and KIM WON-YONG (WON-YONG KIM) *et al., Korean Art Treasures* (1986), a survey of the history of Korean art by Korean experts. Three useful exhibition catalogs are NATIONAL GALLERY OF ART (U.S.), *Masterpieces of Korean Art* (1957); ARTS COUNCIL OF GREAT BRITAIN, *An Exhibition of National Art Treasures of Korea* (1961); and RODERICK WHITFIELD (ed.), *Treasures from Korea: Art Through 5000 Years* (1984), with excellent introductions. Studies of ceramics include ROBERT P. GRIFFING, JR., *The Art of the Korean Potter* (1968), with an excellent introduction; G.ST.G.M. GOMPERTZ, *Korean Celadon, and Other Wares of the Koryŏ Period* (1963); W.B. HONEY, *Korean Pottery* (1947, reissued 1955); and CHEWŎN KIM and G.ST.G.M. GOMPERTZ (eds.), *The Ceramic Art of Korea* (1961). (W.-Y.K.)

Japan: A comprehensive single-volume survey of Japanese art is PENELOPE MASON, *History of Japanese Art* (1993). Serviceable introductions are found in JOAN STANLEY-BAKER, *Japanese Art* (1984); and ROBERT TREAT PAINE and ALEXANDER SOPER, *The Art and Architecture of Japan*, 3rd ed. (1974). Two multivolume series are noteworthy: *The Heibonsha Survey of Japanese Art*, 31 vol. (1971–80); and *Japanese Arts Library* (1977–), are translations and adaptations of Japanese originals, featuring both site-specific and thematic studies. LAURANCE P. ROBERTS, *A Dictionary of Japanese Artists* (1976, reissued 1990), is an indispensable compendium of artists' biographies. *Kodansha Encyclopedia of Japan*, 9 vol. (1983), provides an excellent selection of general and specific articles on Japanese art. SHERMAN E. LEE, *A History of Far Eastern Art*, 5th ed. edited by NAOMI NOBLE RICHARD (1994), relates Japanese art to the wider East Asian context.

The development of Japanese art from Paleolithic times to the 7th century is summarized and related to Chinese and Korean material by GINA L. BARNES, *China, Korea, and Japan: The Rise of Civilization in East Asia* (1993). Other works treating the pre-Buddhist period include RICHARD J. PEARSON *et al., Ancient Japan* (1992); JAPAN, BUNKACHŌ and JAPAN SOCIETY (NEW YORK, N.Y.), *The Rise of a Great Tradition: Japanese Archaeological Ceramics of the Jōmon Through Heian Periods (10,500 BC–AD 1185)* (1990); and J.E. KIDDER, JR., *Japan Before Buddhism*, rev. ed. (1966).

Works that treat the complexity of Buddhist art development include KURATA BUNSAKU (BUNSAKU KURATA), *Hōryū-ji: Temple of the Exalted Law*, trans. from Japanese (1981); JIRŌ SUGIYAMA, *Classic Buddhist Sculpture*, trans. from Japanese and adapted by SAMUEL CROWELL MORSE (1982); YUTAKA MINO *et al., The Great Eastern Temple: Treasures of Japanese Buddhist Art from Tōdai-ji* (1986); NISHIKAWA KYŌTARŌ (KYŌTARŌ NISHIKAWA) and EMILY J. SANO, *The Great Age of Japanese Buddhist Sculpture, AD 600–1300* (1982); JŌJI OKAZAKI, *Pure Land Buddhist Painting*, trans. from Japanese and adapted by ELIZABETH TEN GROTENHUIS (1977); and HISATOYO ISHIDA, *Esoteric Buddhist Painting* (1987; originally published in Japanese, 1969). Other important sources are MIYEKO MURASE, *Emaki: Narrative Scrolls from Japan* (1983); VICTOR HARRIS and KEN MATSUSHIMA, *Kamakura: The Renaissance of Japanese Sculpture, 1186–1333* (1991); and JAN FONTEIN and MONEY HICKMAN, *Zen Painting and Calligraphy* (1970).

The expression of indigenous religion and its relation to Buddhism is summarized in HARUKI KAGEYAMA, *The Arts of Shinto*, trans. from Japanese and adapted by CHRISTINE GUTH (1973). YOSHIAKI SHIMIZU and JOHN M. ROSENFIELD, *Masters of Japanese Calligraphy: 8th–19th Century*, ed. by NAOMI NOBLE RICHARD (1984), discusses the cornerstone of East Asian visual expression.

Daimyo and shogunal patronage from the 15th century is summarized in JAY A. LEVENSON (ed.), *Circa 1492: Art in the Age of Exploration* (1991); and YOSHIAKI SHIMIZU (ed.), *Japan: The Shaping of Daimyo Culture, 1185–1868* (1988). WATANABE AKIYOSHI (AKIYOSHI WATANABE), KANAZAWA HIROSHI (HIROSHI

KANAZAWA), and PAUL VARLEY (H. PAUL VARLEY), *Of Water and Ink* (1986), is a thorough study of the ink monochrome painting style of this period.

Medieval ceramic history is discussed in LOUISE ALLISON CORT, *Shigaraki: Potter's Valley* (1979); JOHANNA BECKER, *Karatsu Ware* (1986); and TSUGIO MIKAMI, *The Art of Japanese Ceramics* (1972; originally published in Japanese, 1968).

The artistically explosive final years of the 16th century are treated in METROPOLITAN MUSEUM OF ART (NEW YORK, N.Y.) and JAPAN, BUNKACHŌ, *Momoyama: Japanese Art in the Age of Grandeur* (1975); TSUNEO TAKEDA, *Kano Eitoku* (1977; originally published in Japanese, 1974); and HAYASHIYA SEIZŌ (SEIZŌ HAYASHIYA), *Chanoyu: Japanese Tea Ceremony* (1979). MOTOO HINAGO, *Japanese Castles*, trans. from Japanese and adapted by WILLIAM H. COALDRAKE (1986); FUMIO HASHIMOTO (ed.), *Architecture in the Shoin Style*, trans. and adapted by H. MACK HORTON (1981; originally published in Japanese, 1972); and MITCHELL BRING and JOSSE WAYEMBERGH, *Japanese Gardens: Design and Meaning* (1981), discuss important innovations in spatial aesthetics.

A broad overview of the artistically eclectic Edo period is found in WILLIAM WATSON (ed.), *The Great Japan Exhibition: Art of the Edo Period, 1600–1868* (1981). The emergence of a distinctive decorative tradition is treated in CAROLYN WHEELWRIGHT (ed.), *Word in Flower: The Visualization of Classical Literature in Seventeenth-Century Japan* (1989); and HOWARD A. LINK, *Exquisite Visions: Rimpa Paintings from Japan* (1980). Chinese-inspired literati painting is summarized in JAMES CAHILL, *Scholar Painters of Japan* (1972); and CALVIN L. FRENCH, *The Poet-Painters: Buson and His Followers* (1974). Other styles of painting are discussed in ST. LOUIS ART MUSEUM and SEATTLE ART MUSEUM, *Ōkyo and the Maruyama-Shijō School of Japanese Painting* (1980); and STEPHEN ADDISS, *The Art of Zen* (1989). MIYAJIMA SHIN'ICHI (SHIN'ICHI MIYAJIMA) and SATŌ YASUHIRO (YASUHIRO SATŌ), *Japanese Ink Painting*, ed. by GEORGE KUWAYAMA (1985), is especially informative on Edo-period "eccentric" painters.

RICHARD LANE, *Images from the Floating World: The Japanese Print* (1978, reissued 1982), combines general essays with a detailed, illustrated dictionary. ROGER S. KEYES, *Japanese Woodblock Prints: A Catalogue of the Mary A. Ainsworth Collection* (1984), provides informative thematic essays interpreting a specific collection. JACK HILLIER, *The Art of the Japanese Book*, 2 vol. (1987), is a comprehensive resource.

Art of the Meiji period is introduced by FREDERICK BAEKELAND and MARTIE W. YOUNG, *Imperial Japan: The Art of the Meiji Era, 1868–1912* (1980). Other works of note include HENRY D. SMITH II, *Kiyochika: Artist of Meiji Japan* (1988); and JULIA MEECH-PEKARIK, *The World of the Meiji Print: Impressions of a New Civilization* (1986). (J.T.U.)

Music. *China:* Western-language sources are listed in FREDRIC LIEBERMAN, *Chinese Music: An Annotated Bibliography*, 2nd, rev. and enlarged ed. (1979). Also of interest are RULAN CHAO PIAN, *Sonq Dynasty Musical Sources and Their Interpretation* (1967); R.H. VAN GULIK, *The Lore of the Chinese Lute*, new ed., rev. (1969); LAURENCE PICKEN (ed.), *Music from the Tang Court*, 5 vol. (1981–90); KENNETH J. DEWOSKIN, *A Song for One or Two: Music and the Concept of Art in Early China* (1982); LIANG MINGYUE (MING-YÜEN LIANG), *Music of the Billion: An Introduction to Chinese Musical Culture* (1985); BELL YUNG, *Cantonese Opera* (1989); COLIN MACKERRAS, *The Rise of the Peking Opera, 1770–1870* (1972), and *The Performing Arts in Contemporary China* (1981); and RICHARD CURT KRAUS, *Pianos and Politics in China: Middle-Class Ambitions and the Struggle over Western Music* (1989).

Korea: The major sources for Korean music study are in Asian languages; some are available in BANG-SONG SONG (trans.), *Source Readings in Korean Music* (1980). General studies are LEE HYE-KU (HYE-GU YI), *Essays on Traditional Korean Music*, trans. and ed. by ROBERT C. PROVINE (1981); and LEE HYE-KU (HYE-GU YI) (compiler and ed.), *Korean Musical Instruments*, trans. by ALAN C. HEYMAN (1982). Older works may be found by consulting BANG-SONG SONG, *An Annotated Bibliography of Korean Music* (1971).

Japan: General introductions are FRANCIS PIGGOTT, *The Music and Musical Instruments of Japan*, 2nd ed. (1909, reprinted 1971); and WILLIAM P. MALM, *Japanese Music and Musical Instruments* (1959, reissued 1990). Special studies are ROBERT GARFIAS, *Music of a Thousand Autumns: The Tōgaku Style of Japanese Court Music* (1975); WILLEM ADRIAANSZ, *The Kumiuta and Danmono Traditions of Japanese Koto Music* (1973); BONNIE C. WADE, *Tegotomono: Music for the Japanese Koto* (1976); WILLIAM P. MALM, *Nagauta: The Heart of Kabuki Music* (1963, reprinted 1973), and *Six Hidden Views of Japanese Music* (1986); and C. ANDREW GERSTLE, KIYOSHI INOBE, and WILLIAM P. MALM, *Theater as Music* (1990), an examination of a bunraku play. (W.P.M.)

Dance and theatre. *General works:* JAMES R. BRANDON (ed.), *The Cambridge Guide to Asian Theatre* (1993); MARTIN BANHAM (ed.), *The Cambridge Guide to World Theatre* (1988); JOEL TRAPIDO (ed.), *An International Dictionary of Theatre Language* (1985); and *Asian Theatre Journal* (semiannual), contain useful entries on East Asian Theatre.

China: Critical studies include A.C. SCOTT, *The Classical Theatre of China* (1957, reprinted 1978), a standard work; the excellent history of the Peking opera by Mackerras cited above in the section on music; LIU WU-CHI (WU-CHI LIU), *An Introduction to Chinese Literature* (1966, reissued 1990), an analysis of individual playwrights and their works; J.I. CRUMP, *Chinese Theater in the Days of Kublai Khan* (1980), a study of Yüan drama; WILT IDEMA and STEPHEN H. WEST, *Chinese Theatre, 1100–1450: A Source Book* (1982), translations of theatre documents; COLIN MACKERRAS (ed.), *Chinese Theater: From Its Origins to the Present Day* (1983), a comprehensive survey; ROBERTA HELMER STALBERG, *China's Puppets* (1984), a well-illustrated introduction; and TAO-CHING HSÜ, *The Chinese Conception of the Theatre* (1985), a compilation of Chinese historical sources.

Korea: Studies of various Korean performing arts are CH'OE SANG-SU (SANG-SU CH'OE), *A Study of the Korean Puppet Play* (1961), a detailed study with illustrations and translations of two play texts; HALLA PAI HUHM, *Kut: Korean Shamanist Rituals* (1980); and KOREAN NATIONAL COMMISSION FOR UNESCO (ed.), *Korean Dance, Theater, and Cinema* (1983).

Japan: Overviews of Japanese performing arts are BENITO ORTOLANI, *The Japanese Theatre* (1990); KAWATAKE TOSHIO (TOSHIO KAWATAKE), *Japan on Stage: Japanese Concepts of Beauty as Shown in the Traditional Theatre* (1990; originally published in Japanese, 1982), on the appeal of traditional performance to Japanese and non-Japanese audiences; JACOB RAZ, *Audience and Actors: A Study of Their Interaction in the Japanese Traditional Theatre* (1983); and YOSHINOBU INOURA and TOSHIO KAWATAKE, *A History of Japanese Theater*, 2 vol. (1971, reissued in 1 vol. as *The Traditional Theater of Japan*, 1981).

MASATARO TOGI, *Gagaku: Court Music and Dance* (1971), provides an overview of this art form's various styles and genres.

Studies of the history and interpretation of nō and *kyōgen* theatre are found in MONICA BETHE and KAREN BRAZELL, *Nō as Performance* (1978); DONALD KEENE and KANEKO HIROSHI (HIROSHI KANEKO), *Nō: The Classical Theatre of Japan*, rev. ed. (1973); KUNIO KOMPARU (KUNIO KONPARU), *The Noh Theater: Principles and Perspectives* (1983; originally published in Japanese, 1980); THOMAS BLENMAN HARE, *Zeami's Style: The Noh Plays of Zeami Motokiyo* (1986); REBECCA TEELE (compiler), *Nō/Kyōgen Masks and Performance* (1984), essays by Japanese artists and Western scholars; and J. THOMAS RIMER and YAMAZAKI MASAKAZU (trans.), *On the Art of the Nō Drama: The Major Treatises of Zeami* (1984).

Historical and interpretive examinations of kabuki include EARLE ERNST, *The Kabuki Theatre* (1956, reissued 1974); MASAKATSU GUNJI and CHIAKI YOSHIDA, *Kabuki*, trans. from Japanese (1969); MASAKATSU GUNJI, *Buyo: The Classical Dance*, trans. from Japanese (1970); and SAMUEL L. LEITER, *Kabuki Encyclopedia* (1979).

Works treating the history and interpretation of bunraku include BARBARA ADACHI, *The Voices and Hands of Bunraku* (1978); DONALD KEENE and KANEKO HIROSHI (HIROSHI KANEKO), *Bunraku: The Art of the Japanese Puppet Theatre*, rev. ed. (1973); C.J. DUNN, *The Early Japanese Puppet Drama* (1966); and C. ANDREW GERSTLE, *Circles of Fantasy: Convention in the Plays of Chikamatsu* (1986).

Modern theatre history and interpretation are explored in ETHAN HOFFMAN *et al.*, *Butoh: Dance of the Dark Soul* (1987); SUSAN BLAKELEY KLEIN, *Ankoku Butō: The Premodern and Postmodern Influences on the Dance of Utter Darkness* (1988); TOYOTAKE KOMIYA, *Japanese Music and Drama in the Meiji Era*, trans. from Japanese (1956, reissued 1969); J. THOMAS RIMER, *Toward a Modern Japanese Theatre* (1974); and TADASHI SUZUKI, *The Way of Acting* (1986; originally published in Japanese, 1984). (J.R.B.)

Eastern Africa

Eastern Africa is a part of sub-Saharan Africa that comprises two traditionally recognized regions: East Africa, made up of Kenya, Tanzania, and Uganda; and the Horn of Africa, made up of Somalia, Djibouti, Eritrea, and Ethiopia. The Horn of Africa, containing such diverse areas as the Ethiopian highlands, the Ogaden desert, and the Eritrean and Somalian coasts, is home to the Amhara, Tigray, Oromo, and Somali peoples, among others. Washed by the Red Sea, the Gulf of Aden, and the Indian Ocean, this region has long been in contact with the Arabian Peninsula and southwest Asia. Islām and Christianity are of ancient standing here, and the people speak Hamito-Semitic tongues related to the languages of North Africa and the Middle East. East Africa, too, has a long history of contact with Arabia, particularly through the island of Zanzibar and the ancient ports of the Swahili coast, but it was through the Bantu kingdoms near Lake Victoria and through the farming and cattle-raising cultures of the Kenyan highlands that this region, early on, showed a much closer affinity with sub-Saharan Africa.

Both regions went through periods of conquest and colonization by European powers, the Horn being controlled by Italy, France, and Great Britain and the East African lands becoming protectorates of Britain and Germany. It is to this era, which finally came to an end in 1977 with the independence of Djibouti, that the seven countries discussed in this article owe their present boundaries.

This article begins with a description of the geography and economy of all of eastern Africa; it then proceeds with sections on the cultures and histories of East Africa and the Horn of Africa. Following these regional treatments are sections on the geography and the history of each country. For more detailed geographic information, see AFRICA. For treatments of regions related to eastern Africa, see CENTRAL AFRICA; NORTH AFRICA; SUDAN, THE; EGYPT; and ARABIA. For artistic expressions, see AFRICAN ARTS. (Ed.)

The article is divided into the following sections:

The region 772
  The land 772
    Relief
    Drainage
    Soils
    Climate
    Plant and animal life
  The economy 776
    Agriculture
    Forestry
    Fishing
    Resources
    Commerce and industry
  The people 778
    East Africa
    The Horn of Africa
  History 784
    East Africa
    The Horn of Africa
The countries of East Africa 795
  Kenya 795
    Physical and human geography
    History
  Tanzania 803
    Physical and human geography
    History
  Uganda 810
    Physical and human geography
    History
The countries of the Horn of Africa 817
  Djibouti 817
    Physical and human geography
    History
  Eritrea 819
    Physical and human geography
    History
  Ethiopia 823
    Physical and human geography
    History
  Somalia 831
    Physical and human geography
    History
Bibliography 836

THE REGION

The land

RELIEF

The physical basis of eastern Africa is a platform of ancient resistant rocks that has been contorted and inset with granites but worn down by prolonged erosion to extensive plains. Its present outlines derive from the splitting apart of the ancient supercontinent of Gondwanaland, of which Africa forms a part. In eastern Africa the straight coastlines of Eritrea and northern Somalia were created by the drifting away of the Arabian Peninsula, which opened up the Red Sea and the Gulf of Aden, and the smooth shorelines and deep waters along the eastern coast mark the departure of India and Madagascar. Too rigid for folding to take place, the platform on which eastern Africa rests has been buckled by subterranean forces into broad basin-and-swell structures hundreds of miles across. Associated with these tensional forces, extensive faulting has raised and lowered vast blocks of land, leaving prominent escarpments between them, and extruded lavas have formed elevated plateaus and have spread across the plains as well as forming numerous volcanoes. The most striking of these features is the East African Rift System, of which the main branch, known as the Eastern Rift Valley or Great Rift Valley, extends from the junction of the Red Sea and the Gulf of Aden, crosses the summit of two centres of uplift in Ethiopia and Kenya, and enters northern Tanzania, where it largely disappears only to reappear in the south of that country in the Lake Nyasa trough (Lake Nyasa is also known as Lake Malawi). The Western Rift Valley curves along the western border of Uganda and Tanzania, where it is marked by Lakes Albert and Tanganyika, and is aligned through the Lake Rukwa trough with the head of Lake Nyasa. Although not entirely continuous or uniform, the rift valleys are typically some 35 miles (60 kilometres) across and, where they cut through highland, may have inward-facing scarps of 1,500 to 3,000 feet (500 to 1,000 metres) in elevation. The two most striking highlands, found in Ethiopia and Kenya, are formed of lava flows piled on top of areas of uplift on either side of the Great Rift.

The East African Rift System

These fundamental geologic factors are reflected in the major physiographic regions of eastern Africa. The Ethiopian highlands, for example, are formed from lava flows that have created extensive plateaus at elevations of 6,500 to 10,000 feet. The plateaus are separated by deep, river-worn gorges and are marked by isolated summits rising to over 12,000 feet. The northern end of the Rift Valley

is a region of confused relief, characterized by downfaulting to below sea level in the Kobar Sink and by active volcanoes and hot mineral springs. The Kenyan highlands are constructed by lava flows piled upon a broad, uplifted dome that is dissected by the Great Rift Valley. There the shoulders of the Rift highlands rise to nearly 12,000 feet, but of greater height are giant extinct volcanoes on the outer edge of the volcanic province—Mounts Elgon and Kenya and Kilimanjaro, the latter, at 19,340 feet (5,895 metres), the highest mountain in Africa. In southern Tanzania the continuation of the Great Rift Valley is bordered by the Southern and Nyasa highlands, which overlook Lake Nyasa; and the Western Rift is bordered by the Ufipa Plateau, which lies above Lake Rukwa. In Uganda the Western Rift Valley is flanked by high ground in Kigezi and Karagwe and by the upfaulted block of the Ruwenzori Range.

Between the arms of the two Rift valleys lies the Central Plateau, an extensive, eroded surface comprising most of Uganda and western Tanzania. Lying mostly at 3,000 to 4,500 feet, it is a major example of a peneplain created by long periods of erosion but bearing isolated ridges and hill masses of more resistant material called inselbergs. East of the Great Rift, the surface is further diversified by faulting and then gives way to a coastal zone of sedimentary stratified rocks; this creates a gently varied relief of plateaus, escarpments, and riverine plains. In the Tana River basin of eastern Kenya and in most of Somalia, the original land surface has sunk thousands of feet below sea level; this has been covered by more recent sediments, which have resulted in extensive and very complete plains.

DRAINAGE

Inland basins

Rivers and lakes. The Great Rift Valley is the centre of a remarkable line of inland drainage basins; radiating outward from its bordering highlands, other waters drain to the Indian and Atlantic oceans and to the Mediterranean Sea. The area of inland drainage extends from Lake Abaya in southern Ethiopia through Lake Rudolf (or Lake Turkana) in Kenya to the strongly alkaline lakes of Natron, Manyara, and Eyasi in northern Tanzania. Lake Rukwa is the centre of a separate basin of inland drainage. There is little drainage in the arid coastlands of the Red Sea and the Gulf of Aden, but the strongly seasonal Shebele (or Shabeelle) and Jubba rivers manage to carry runoff from the summer rains of Ethiopia across Somalia to the Indian Ocean. The Tana and Athi-Galana systems from the Kenyan highlands are more reliable, as are those of eastern Tanzania, notably the extensive Rufiji–Kilombero–Great Ruaha system.

The Nile has its headwaters in the eastern African highlands and plateaus, forcing Egypt to maintain an interest in dams built in Ethiopia and Uganda. (Actually, it is the Blue Nile, or Abay, River and the Atbara and Sobat rivers that bring seasonal floods to The Sudan and Egypt, while the more regular flow of the White Nile is derived from Lake Victoria.) Another great river, the Congo, receives contributions from the southern portion of the Central Plateau through the Malagarasi River, which debouches into Lake Tanganyika. This lake is some 400 miles long but has an average width of only 30 miles. Also, although it has a surface elevation of some 2,500 feet, its bottom reaches to about 2,200 feet below sea level.

Groundwater. The seasonal nature and general scarcity of rainfall over much of eastern Africa makes groundwater supplies especially significant. The volcanic strata of the Rift highlands are particularly useful in storing water, releasing it into springs and rivers or providing good well sites. But water does not so easily penetrate the hard rocks of the underlying platform, where groundwater tends to be limited to small and localized pockets. The layered sediments of Somalia and the coastal plains absorb rainfall, but it is liable to be lost to great depths and, if reached by boreholes, may prove saline.

SOILS

Soils vary greatly, but a broad pattern can be discerned in relation to climate (especially rainfall), to drainage, and to parent material.

Soils of the drier regions. The arid and semiarid zones of the Eritrean coast, Somalia, northern Kenya, and the Ogaden region of Ethiopia are mostly covered with shallow soils, often stony and little-weathered. They include almost bare lavas near the Rift Valley and calcareous clay loams over sedimentary limestones to the east. Residual gypsum is common in northern Somalia. Less arid areas are characterized by intermediate or dryland soils that are more thoroughly, but not deeply, weathered; these reddish soils are rich in iron, which is often found as granules on the surface or in a buried layer, and they are of average agricultural value, being useful under irrigation. Less fertile are the old and greatly weathered soils of much of Tanzania, which form a recurrent topographic sequence, or catena, according to whether they are located on a level plain, on a gentle slope, or in a broad, shallow drainage channel at the foot of a slope. The upper plateau soils tend to be deeply weathered and leached sandy loams. On the slopes, soil movement exposes less weathered material, and fertility and drainage are more favourable for agriculture. The drainage channels are frequently waterlogged, giving rise to mottled sandy loams or to a black clay that can be richer in minerals and nutrients but difficult to cultivate. Over extensive areas of the plateaus, the residual iron content of the soil is enough to form a crust of ironstone or laterite.

Infertile soils

Soils of the wetter regions. Highland soils form a distinctive category because of the climate but also because so many of them are derived from volcanic material. The bright red, rich clay loams of Kenya and similar volcanic uplands result from deep weathering under ample rainfall, yielding a highly fertile basis for agriculture. At cooler elevations above about 6,500 feet, soil colour changes to a deep brown, but fertility, as on the Ethiopian Plateau, remains high. Poorly drained lava surfaces can weather into plains of black cracking clay (also called vertisol or "black cotton" soil), a poor foundation for buildings or roadworks. Soil erosion is particularly serious on steep slopes of the highlands of Eritrea, Ethiopia, and Kenya, where clearance of the prized soils for cultivation leads to silt-laden rivers, gullied landscapes, and loss of topsoil.

CLIMATE

Straddling the Equator from latitudes 18° N to 18° S, eastern Africa's climate is dominated by its tropical location and by a great range of elevation. Average temperatures are reduced by the high average elevation, but only on the highest mountains is the temperature low enough to restrict the growth of vegetation. It is the amount and seasonal duration of rainfall that distinguishes most climatic regions. As the sun moves into either the northern or southern tropic, so converging air flows, which are uplifted as they meet at a zone of low pressure called the intertropical convergence zone, bring intense summer rains; these are followed by a winter dry season as the sun shifts to the other tropic. Thus, northern Uganda and central and southern Tanzania receive 20 to 48 inches (500 to 1,200 millimetres) of rainfall in a five- to eight-month season. Around Lake Victoria on the Equator, more continuous rains follow from two seasons of overhead sun and from the local effects of the vast surface of the lake. Arid Somalia and northeastern Kenya are anomalous in these latitudes, with rainfall less than 10 inches per year. There, in the northern summer, airflow diverges toward the low pressure of the Indian Ocean monsoon system, resulting in a gently subsiding atmosphere rather than the uplift needed to generate precipitation. In winter a contrary outflow from Southwest Asia brings little moisture, but one along the Red Sea brings winter rain to dry coastal Eritrea.

Effects of elevation on climate

The major highlands are sufficiently extensive to form a major exception to these patterns. Even at the Equator, a reduction in temperature at elevations above 5,400 feet creates climates outside the tropical category, with important implications for agricultural ecology and health. The highest mountain summits rate as alpine, with glaciers present on Kilimanjaro and other peaks. Even lower relief features are sufficient to generate locally enhanced precipitation, and coastal locations in Tanzania and southern Kenya also experience locally high rainfalls.


Port Sudan
SAUDI ARABIA
RED SEA
DAHLAK ARCHIPELAGO
Khartoum
ERITREA
Asmara
San'a'
YEMEN
Aden
Gulf of Aden
DJIBOUTI
Djibouti
Cape Guardafuy
Boosaaso
Point Xaafuun
THE SUDAN
DANAKIL DEPRESSION
Mekele
Gonder
Bahir Dar
ETHIOPIAN PLATEAU
Addis Ababa
Nazret
Dire Dawa
Hargeysa
Berbera
DHAROOR VALLEY
KARKAR MOUNTAINS
NUGAALEED VALLEY
HAWD
OGADEN
ETHIOPIA
SOMALIA
Jima
Shashemene
Kibre Mengist
Hobyo
Beledweyne
Mogadishu
Merka
Baraawe
Kismaayo
Jawhar
Buulobarde
Baydhabo
Bardheere
Marka
Ismaamie
Equator
INDIAN OCEAN
LAG BADANA NATIONAL PARK
DEMOCRATIC REPUBLIC OF THE CONGO
UGANDA
KENYA
Kampala
Jinja
Kisumu
Eldoret
Nakuru
Meru
Nairobi
Machakos
Kisii
Kigali
RWANDA
BURUNDI
Bujumbura
Mwanza
Kigoma
Tabora
TANZANIA
Tanga
Mombasa
Zanzibar
ZANZIBAR
Dar es Salaam
Morogoro
Mbeya
ZAMBIA
MALAWI
MOZAMBIQUE
Lilongwe
COMOROS
Moroni
Tunduru
Mtwara
Nkwalo
Lindi
Kondoa
Dodoma
Iringa
Arusha
Moshi
Kilimanjaro
Mount Kenya
Lake Victoria
Lake Turkana
Lake Rudolf
Lake Tana
Lake Nasser
Lake Albert
Lake Edward
Lake Kivu
Lake Tanganyika
Lake Malawi
Lake Nyasa
Lake Rukwa
Lake Natron
Lake Manyara
Lake Eyasi
Lake Rukwa
Nile
Blue Nile
Atbara
Sobat
Omo
Tekeze
Awash
Shebelle
Juba
Webi Shabeelle
Ogaden
Jubba
Tana
Athi
Ewaso Ngiro
Malagarasi
Ruvuma
Rufiji
Kagera
Zambezi
Shire
Luangwa
Luapula
Kafue
Limpopo
Okavango
Cuando
Chobe
Linyanti
Kwando
Zambezi
Etosha Pan
Makgadikgadi Pans
Namib Desert
Kalahari Desert
Great Rift Valley
Ethiopian Highlands
Mount Elgon
Mount Kilimanjaro
Mount Meru
Mount Kenya
Aberdare Range
Mau Escarpment
Mount Stanley
Mount Gahinga
Mount Muhavura
Mount Karisimbi
Mount Nyiragongo
Mount Mikeno
Mount Ruwenzori
Mount Wuhlu
Mount Moroto
Mount Otoke
Mount Debasien
Mount Kadam
Mount Lolmalasin
Mount Suswa
Mount Longonot
Mount Menengai
Mount Elgon
Mount Kadam
Mount Lo

MAP INDEX

Cities and towns

Addis Ababa (Adis Abeba) .... 9 02 N 38 42 E
Adigrat ..... 14 17 N 39 28 E
Adola ..... 11 48 N 41 42 E
Adwa (Adowa or Aduwa) ..... 14 10 N 38 54 E
Agaro ..... 7 51 N 36 39 E
Akaki ..... 9 05 N 39 00 E
Akordat ..... 15 33 N 37 53 E
Aksum (Axum) ..... 14 08 N 38 43 E
Alamata ..... 12 25 N 39 33 E
Ali Sabih ..... 11 10 N 42 42 E
Arba Minch (Arba Merch) .... 6 02 N 37 33 E
Arusha ..... 3 22 S 36 41 E
Aseb (Assab) .... 13 00 N 42 44 E
Asela (Asala) .... 7 57 N 39 08 E
Asmera (Asmara) .... 15 20 N 38 56 E
Assab, see Aseb
Awasa ..... 7 03 N 38 28 E
Axum, see Aksum
Baardheere (Bardera) ..... 2 20 N 42 17 E
Bagamoyo ..... 6 26 S 38 54 E
Bahir Dar ..... 11 36 N 37 23 E
Baidoa, see Baydhabo
Barawe (Brava) .. 1 06 N 44 03 E
Bardera, see Baardheere
Baydhabo (Baidoa) ..... 3 07 N 43 39 E
Beledweyne (Belet Uen) ..... 4 45 N 45 12 E
Bender Cassim, see Boosaaso
Berbera ..... 10 25 N 45 02 E
Boosaaso (Bender Cassim) ..... 11 17 N 49 11 E
Brava, see Baraawe
Bukoba ..... 1 20 S 31 49 E
Bulo Burti, see Buulobarde
Bungoma ..... 0 34 N 34 34 E
Burao (Burco) .... 9 31 N 45 32 E
Busia ..... 0 28 N 34 06 E
Buulobarde (Bulo Burti) ..... 3 51 N 45 34 E
Ceerigaabo (Erigavo) ..... 10 37 N 47 22 E
Chake Chake .... 5 15 S 39 46 E
Chisimayu, see Kismaayo
Dar es Salaam .... 6 48 S 39 17 E
Debre Birhan (Debre Berhan) .... 9 41 N 39 32 E
Debre Markos .... 10 21 N 37 44 E
Debre Tabor .... 11 51 N 38 01 E
Debre Zeyit .... 8 45 N 38 59 E
Dembidollo ..... 8 32 N 38 48 E
Dese (Dase) .... 11 08 N 39 38 E
Dikhil ..... 11 06 N 42 23 E
Dila ..... 6 25 N 38 19 E
Dire Dawa ..... 9 35 N 41 52 E
Djibouti ..... 11 36 N 43 09 E
Dodoma ..... 6 11 S 35 45 E
Eldoret ..... 0 31 N 35 17 E
Embu ..... 0 32 S 37 27 E
Entebbe ..... 0 04 N 32 28 E
Erigavo, see Ceerigaabo
Eyl ..... 7 59 N 49 49 E
Fiche ..... 9 48 N 38 44 E
Finchaa ..... 9 33 N 37 21 E
Garissa ..... 0 28 S 39 38 E
Ghion, see Giyon
Giamama, see Jamaame
Gohar, see Jawhar
Giyon (Ghion) .... 8 32 N 37 59 E
Goba ..... 7 01 N 39 59 E
Gonder ..... 12 36 N 37 28 E
Gore ..... 8 09 N 35 32 E
Gulu ..... 2 47 N 32 18 E
Hagere Hiywet (Hagere Hiwet) .. 8 59 N 37 51 E
Harer (Harar) .... 9 19 N 42 07 E
Hargeysa ..... 9 35 N 44 04 E
Hobyo (Obbia) .... 5 21 N 48 32 E
Hosaina (Hosana) ..... 7 35 N 37 53 E
Ifakara ..... 8 08 S 36 41 E
Iringa ..... 7 46 S 35 42 E
Isiolo ..... 0 21 N 37 35 E
Jamaame (Giamama or Jamaame or Margherita) ..... 0 04 N 42 45 E
Jawhar (Gohar) .... 2 46 N 45 31 E
Jijiga ..... 9 21 N 42 48 E
Jima (Jimma) .... 7 40 N 36 50 E
Jinja ..... 0 26 N 33 12 E
Kabale ..... 1 15 S 29 59 E
Kabarole ..... 0 39 N 30 16 E
Kakamega ..... 0 17 N 34 45 E
Kampala ..... 0 19 N 32 35 E
Kembolcha (Kombolcha) .... 11 05 N 39 44 E
Keren ..... 15 47 N 38 28 E
Kericho ..... 0 22 S 35 17 E
Kibre Mengist ..... 5 53 N 38 59 E
Kigoma ..... 4 52 S 29 38 E
Kilosa ..... 6 50 S 36 59 E
Kisii ..... 0 41 S 34 46 E
Kismaayo (Chisimayu) ..... 0 22 S 42 32 E
Kisumu ..... 0 06 S 34 45 E
Kitale ..... 1 01 N 34 00 E
Kombolcha, see Kembolcha
Korogwe ..... 5 09 S 38 29 E
Lalibela ..... 12 02 N 39 02 E
Lamu ..... 2 16 S 40 54 E
Lindi ..... 10 00 S 39 43 E
Lodwar ..... 3 07 N 35 36 E
Machakos ..... 1 31 S 37 16 E
Malindi ..... 3 13 S 40 07 E
Mandera ..... 3 56 N 41 52 E
Maralal ..... 1 06 N 36 42 E
Margherita, see Jamaame
Marka (Merca) .... 1 43 N 44 53 E
Marsabit ..... 2 20 N 37 59 E
Masaka ..... 0 20 S 31 44 E
Massawa, see Mitsiwa
Maychew ..... 12 47 N 39 32 E
Mbale ..... 1 05 N 34 10 E
Mbarara ..... 0 37 S 30 39 E
Mbeya ..... 8 54 S 33 27 E
Meiso ..... 13 30 N 39 28 E
Merca, see Marka
Meru ..... 0 03 N 37 39 E
Metu ..... 8 18 N 35 35 E
Mitsiwa (Massawa) ..... 15 36 N 39 28 E
Mkoani ..... 5 22 S 39 39 E
Mogadishu (Mogadiscio or Muqdisho) .... 2 04 N 45 22 E
Mojo ..... 8 36 N 39 07 E
Mombasa ..... 4 03 S 39 40 E
Morogoro ..... 6 49 S 37 40 E
Moshi ..... 3 21 S 37 20 E
Mpwapwa ..... 6 21 S 36 29 E
Mtwara ..... 10 16 S 40 11 E
Muqdisho, see Mogadishu
Murang'a ..... 0 43 S 37 09 E
Musoma ..... 1 30 S 33 48 E
Mwadui ..... 3 33 S 33 36 E
Mwanza ..... 2 31 S 32 54 E
Nairobi ..... 1 17 S 36 49 E
Naivasha ..... 0 43 S 36 26 E
Nakfa ..... 16 40 N 38 29 E
Nakuru ..... 0 17 S 36 04 E
Nanyuki ..... 0 01 N 37 04 E
Nazret ..... 8 33 N 39 16 E
Nekemte ..... 9 05 N 36 33 E
Newale ..... 10 56 S 39 18 E
Nyahururu Falls .... 0 02 N 36 22 E
Nyeri ..... 0 25 S 36 57 E
Obbia, see Hobyo
Pangani ..... 9 32 S 35 31 E
Seylac (Zeila) .... 11 21 N 43 29 E
Shashemene .... 7 12 N 38 36 E
Shinyanga ..... 3 40 S 33 26 E
Singida ..... 4 49 S 34 45 E
Sodo ..... 6 54 N 37 45 E
Songea ..... 10 41 S 35 39 E
Soroti ..... 1 43 N 33 37 E
Sumbawanga .... 7 58 S 31 37 E
Tabora ..... 5 01 S 32 48 E
Tadjoura ..... 11 47 N 42 53 E
Tanga ..... 5 04 S 39 06 E
Teseney ..... 15 08 N 36 40 E
Thika ..... 1 03 S 37 05 E
Tororo ..... 0 42 N 34 11 E
Tunduru ..... 11 07 S 37 21 E
Voi ..... 3 23 S 38 34 E
Wajir ..... 1 45 N 40 04 E
Weldya ..... 11 50 N 39 41 E
Wenji ..... 8 27 N 39 17 E
Wete ..... 5 04 S 39 43 E
Xaafuun ..... 10 25 N 51 16 E
Yirga Alem ..... 6 45 N 38 25 E
Zanzibar ..... 6 10 S 39 11 E
Zeila, see Seylac

Physical features and points of interest

Abay, see Blue Nile
Abaya, Lake ..... 6 20 N 37 50 E
Abe, Lake ..... 11 10 N 41 47 E
Aberdare Range .. 0 25 S 36 38 E
Achwa (Aswa), *river* ..... 3 43 N 31 55 E
Aden, Gulf of .... 12 00 N 48 00 E
Ahmar Mountains .. 9 23 N 41 13 E
Albert, Lake ..... 1 40 N 31 00 E
Anseba, *river* .... 17 03 N 37 24 E
Aswa, see Achwa
Atbara (Atbarah), *river* ..... 17 40 N 33 58 E
Athi, *river* ..... 2 59 S 38 31 E
Awash (Hawash), *river* ..... 11 35 N 41 38 E
Bale Mountains National Park .... 6 45 N 39 45 E
Baraka, *river* ..... 18 13 N 37 35 E
Baringo, Lake .... 0 38 N 36 05 E
Baro, *river* ..... 8 23 N 33 11 E
Batu, Mount .... 6 55 N 39 44 E
Benadir, *region* .. 2 40 N 45 45 E
Blue Nile (Abay or Al-Azraq), *river* .. 15 38 N 32 31 E
Bokora Corridor Game Reserve .. 2 30 N 34 05 E
Bor, *river* ..... 1 18 N 40 40 E
Cal Madow Mountains ..... 11 00 N 48 30 E
Chalbi Desert .... 3 00 N 37 20 E
Chamo, Lake ..... 5 50 N 37 33 E
Chew Bahir (Stefanie), Lake .. 4 38 N 36 50 E
Choke Mountains .. 10 45 N 37 35 E
Dahlak Archipelago ..... 15 50 N 40 12 E
Dawa (Daawo or Daua), *river* .... 4 11 N 42 05 E
Denakil, *region* .... 13 00 N 41 00 E
Dera, *river* ..... 0 15 N 42 17 E
Dharoor (Daror) Valley ..... 10 20 N 50 20 E
Dopeth, *river* .... 2 40 N 34 02 E
East African Rift System, see Eastern Rift Valley and Western Rift Valley
Eastern (Great) Rift Valley ..... 5 00 N 37 00 E
Edward, Lake .... 0 25 S 29 30 E
Elgon, Mount .... 1 08 N 34 33 E
Ethiopian Plateau .. 10 00 N 38 10 E
Eyasi, Lake ..... 3 40 S 35 05 E
Eyl, *river* ..... 7 58 N 49 52 E
Fillo, Mount ..... 7 22 N 39 21 E
Galana, *river* .... 3 09 S 40 08 E
Galgodon (Ogo) Mountains ..... 9 50 N 45 20 E
Gambella National Park .... 8 00 N 34 10 E
Gara Muleta, Mount ..... 9 15 N 41 44 E
Gash (Al-Qash), *river* ..... 16 48 N 35 51 E
Genale, see Jubba
George, Lake .... 0 00 N 30 12 E
Great Rift Valley, see Eastern Rift Valley
Great Ruaha, *river* ..... 7 56 S 37 52 E
Guban Plain ..... 10 15 N 44 26 E
Guge, Mount .... 6 12 N 37 30 E
Guna, Mount .... 11 43 N 38 14 E
Gwardafuy (Guardafui), Cape ..... 11 49 N 51 15 E
Hawash, see Awash
Hawd (Haud), *region* ..... 8 00 N 46 30 E
Indian Ocean ..... 2 30 S 47 50 E
Jaceyl, *river* .... 10 22 N 50 58 E
Jubba (Genale), *river* ..... 0 15 S 42 38 E
Kalabega, see Murchison Falls National Park
Kagera, *river* .... 0 57 S 31 47 E
Kalambo Falls .... 8 36 S 31 14 E
Kar Kar Mountains ..... 9 57 N 49 20 E
Kasigau Hill ..... 3 50 S 38 40 E
Kenya (Kirinyaga), Mount ..... 0 10 S 37 20 E
Keno, *river* ..... 2 59 N 36 07 E
Kilimanjaro, *volcano* ..... 3 04 S 37 22 E
Kilombero, *river* .. 8 31 S 37 22 E
Kirinyaga, see Kenya, Mount
Kizigo Game Reserve ..... 6 20 S 34 00 E
Kobar Sink, *depression* ..... 13 35 N 40 50 E
Kora National Park ..... 0 20 S 38 45 E
Kyoga, Lake ..... 1 30 N 33 00 E
Lag Badana National Park .... 1 20 S 41 35 E
Mafia Island ..... 7 50 S 39 50 E
Magadi, Lake, *lake bed* ..... 1 52 S 36 17 E
Mago National Park ..... 5 30 N 36 15 E
Malagarasi, *river* .. 5 12 S 29 47 E
Malawi, see Nyasa, Lake
Manyara, Lake ... 3 35 S 35 50 E
Mara, *river* ..... 1 31 S 33 56 E
Margherita (Marguerite) Peak ..... 0 22 N 29 51 E
Marsabit National Reserve ..... 2 25 N 37 58 E
Masai Steppe .... 4 45 S 37 00 E
Mau Escarpment .. 0 40 S 36 02 E
Mbeya Range .... 8 50 S 33 20 E
Mereb, *river* .... 15 05 N 37 10 E
Mikumi National Park ..... 7 15 S 37 05 E
Moroto, Mount .... 2 32 N 34 46 E
Morungole, Mount ..... 3 49 N 34 02 E
Mousa, Mount, *volcano* ..... 12 28 N 42 24 E
Moyowosi Game Reserve ..... 3 50 S 31 30 E
Muhavura, *volcano* ..... 1 23 S 29 40 E
Muhuwesi, *river* ... 11 16 S 37 58 E
Murchison Falls .... 2 17 N 31 41 E
Murchison Falls (Kabalega) National Park .... 2 15 N 31 50 E
Muzizi, *river* .... 1 03 N 30 33 E
Natron, Lake ..... 2 25 S 36 00 E
Ngorongoro Conservation Area ..... 3 00 S 35 25 E
Ngorongoro Crater ..... 3 10 S 35 35 E
Nguru Mountains .. 6 00 S 37 30 E
Nugaaleed (Nogal) Valley ... 8 35 N 48 35 E
Nyasa (Malawi), Lake ..... 12 00 S 34 30 E
Ogaden, *region* ... 8 00 N 44 00 E
Ogo, see Galgodon Mountains
Olduvai Gorge, *historical site* .... 2 58 S 35 22 E
Omo, *river* ..... 4 31 N 35 59 E
Omo National Park ..... 6 00 N 35 50 E
Owen Falls Dam .. 0 27 N 33 11 E
Pager, *river* ..... 3 09 N 32 30 E
Pangani, *river* .... 5 26 S 38 58 E

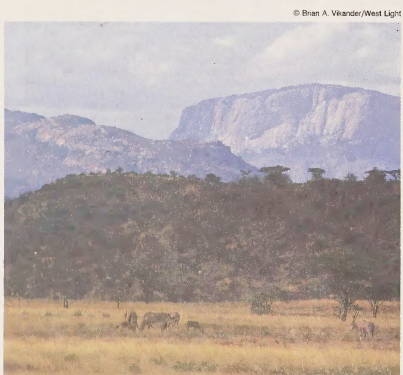
Pare Mountains . . . 4 00 S 37 45 E
Pemba, *island* . . . . 5 10 S 39 48 E
Queen Elizabeth (Ruwenzori) National Park . . . 0 15 S 30 00 E
Ras Dejen (Ras Dashen), Mount . . . . . 13 16 N 38 24 E
Red Sea . . . . . 17 00 N 41 00 E
Ruaha National Park . . . . . 7 30 S 34 30 E
Rubeho Mountains . . . . . 6 55 S 36 30 E
Rudolf (Turkana), Lake . . . . . 3 30 N 36 00 E
Rufiji, *river* . . . . . 8 00 S 39 20 E
Rukwa, Lake . . . . . 8 00 S 32 25 E
Rungwa Game Reserve . . . . . 7 15 S 34 30 E
Rungwe Mountain . . . . . 9 08 S 33 40 E
Ruvuma, *river* . . . . 10 29 S 40 28 E
Ruwenzori, see Queen Elizabeth National Park
Ruwenzori Range . . 0 23 N 29 54 E
Selous Game Reserve . . . . . 9 00 S 37 30 E
Serengeti National Park . . . . 2 20 S 34 50 E
Serengeti Plain . . . . 2 50 S 35 00 E
Shabeelle (Shebele), *river* . . 0 12 N 42 45 E
Shala, Lake . . . . . 7 29 N 38 32 E
Shebele, see Shabeelle
Sibiloi National Park . . . . . 4 00 N 36 15 E
Simen Mountains . . 13 20 N 38 20 E
Soira, Mount . . . . . 14 45 N 39 32 E
South Kitui National Reserve . 1 50 S 38 45 E
Southern Highlands . . . . . 8 15 S 35 15 E
Stefanie, see Chew Bahir, Lake
Surud Cad, Mount . . . . . 10 44 N 47 14 E
Tadjoura, Gulf of . . 11 40 N 43 00 E
Taita Hills . . . . . 3 25 S 38 20 E
Tana, *river* . . . . . 2 32 S 40 31 E
Tana, Lake . . . . . 12 00 N 37 20 E
Tanganyika, Lake . . 6 00 S 29 30 E
Tekeze (Satit), *river* . . . . . 14 20 N 35 50 E
Tsavo National Park . . . . . 3 00 S 38 40 E
Turkana, see Rudolf, Lake
Turkwel, *river* . . . . 3 06 N 36 06 E
Ufipa Plateau . . . . 8 00 S 31 35 E
Usambara Mountains . . . . . 4 45 S 38 30 E
Victoria, Lake . . . . 1 00 S 33 00 E
Victoria Nile, *river* . . . . . 2 14 N 31 26 E
Wami, *river* . . . . . 6 08 S 38 49 E
Western Rift Valley . . . . . 1 00 S 30 00 E
Weyb, *river* . . . . . 4 17 N 42 02 E
Winam Bay . . . . . 0 15 S 34 35 E
Xaafuun, Point . . . . 10 27 N 51 24 E
Yangudi Rassa National Park . . . 10 45 N 41 05 E
Zanzibar, *island* . . 6 10 S 39 20 E
Ziway, Lake . . . . . 8 00 N 38 50 E

With rainfall so dependent on airflow, fluctuations in the large-scale dynamic systems can move the boundary of adequate moisture hundreds of miles, bringing drought to such areas as highland Eritrea, Tigray in Ethiopia, Machakos in Kenya, and Dodoma in Tanzania.

PLANT AND ANIMAL LIFE

Plants. Vegetation types mirror the rainfall zones, starting with scanty plant cover in the arid and semiarid areas, where infrequent succulents and stunted thornbushes survive the dry seasons and where the brief periods of rain bring short-lived ephemeral herbs and annual grasses. In more moist areas, with seasonal rainfall over 12 inches but with a pronounced dry season, the vegetation, often termed savanna, may be divided into three major physiognomic types: bushland, woodland, and wooded grassland. Bushland, characterizing the drier areas, forms a cover of small trees branching from the base with little grass between. Where this cover is dense, impenetrable thickets may be formed. Woodland is a mantle of deciduous trees whose crowns more or less touch to form a light but almost continuous canopy over a layer of grasses, herbs, and small shrubs. Its greatest extent is over the plateau of central and southern Tanzania, where rainfall totals are 32 to 48 inches per year but where there is a severe dry season of up to six months. Wooded grassland is an open mixture of trees and shrubs standing among a good growth of grass but not forming a canopy over it. In such areas the dry season seldom lasts more than three months, and this type of vegetation may actually be derived from forest cleared by human activities.

© Brian A. Vikander/West Light



Escarpments of the Great Rift Valley rising above the plain north of Samburu Game Preserve, central Kenya. Beisa oryx graze in the foreground.

True forest in areas of low and middle elevation is not common in eastern Africa; where it formerly existed, it has in many places been cleared, as in southern Uganda and along parts of the Kenyan and Tanzanian coasts. Better preserved than these are the montane forests of the Ethiopian and Kenyan highlands. At altitudes above the timberline are heather and moorlands of Afro-Alpine vegetation.

Natural grasslands are rare and are usually caused by special circumstances, such as a high water table or cracking clays, which are disruptive to the roots of larger plants. Other vegetation types not primarily related to climate are freshwater papyrus swamps, which are locally important in southern Uganda, and mangrove forests. Modification by human activity includes deforestation, but there is also a less conspicuous diminution of plant cover and degradation of the savanna areas.

Animals. Animal life also has diminished in response to human pressures, but East Africa in particular remains justly famous for its wildlife, which includes spectacular assemblages of big game. It seems likely that these survived into the 20th century because of a low human population; also, the traditional pastoral cultures of East Africa were tolerant of the competition of wild herbivores, each tending to have its own preferred habitat. The Serengeti Plain of Tanzania still supports large migratory herds of zebra, wildebeest, antelope, and gazelle as well as the lions, cheetahs, and wild dogs that prey on them. Elephants and rhinoceroses favour more wooded areas, and in the forests are buffaloes, bushbucks, rare chimpanzees, and leopards. This pattern, too, has been affected by the spread of human settlement, which has forced animals into environments less able to support them. For example, the extension of agriculture into the wooded grasslands has confined the elephant to drier bushlands, where its browsing causes more havoc. Such disruption has been countered by restrictions on hunting and by the creation of nature reserves and national parks, of which the most famous are Serengeti in Tanzania, Amboseli and Tsavo in Kenya, and Murchison Falls (Kabalega) and Queen Elizabeth (Ruwenzori) in Uganda.

Wildlife of East Africa

Birdlife is abundant, large species including the ostrich of the plains and the flamingo and pelican of the Rift Valley lakes.

An important factor that has the effect of neutralizing human pressures and keeping land available for wildlife is infestation with the tsetse fly, which covers more than 40 percent of Kenya, Uganda, and Tanzania. Other significant insect pests include the locust, malaria-carrying mosquitoes, and the *Simulium* fly, the carrier of onchocerciasis, or river blindness.

The economy

The economies of the eastern African countries are closely related to their natural resources. The great majority of the population is directly dependent upon agriculture or pastoralism for its livelihood, and most exports are of primary agricultural products. A substantial portion of agriculture is on the subsistence level—that is, the raising of foodstuffs necessary for maintaining a livelihood, with no planned surplus left over for trade.

AGRICULTURE

Intensive cultivation. Rainfall is the dominant influence on agricultural output and, hence, on the densities of population. This basic resource varies greatly among the countries of eastern Africa. Without irrigation, arable agriculture requires a reliable annual rainfall of over 30 inches (750 millimetres). In four years out of five, this total may be expected by 78 percent of Uganda and 51 percent of Tanzania but only by 15 percent of Kenya. (The proportion of Somalia that receives this total is negligible, and in Ethiopia the range of elevations makes such totals not significant.) A large area of high-intensity agriculture is shared by the three East African countries in the Lake Victoria basin, especially in an arc from western Kenya through Buganda to Bukoba in Tanzania. Food crops here include the banana, sweet potato, taro, and yam, with Robusta coffee and cotton important cash crops. Along the East African coast, between Malindi and Dar es Salaam and including Zanzibar and Pemba, is another closely settled zone with an economy and culture enriched by a thousand years of trading with Arabia, the Persian Gulf, and the Indian subcontinent.

Highland agriculture

Other intensively cultivated areas are in the uplands and mountains, where precipitation, increased by the raised landforms, is made more available for plant growth because the cooler temperatures reduce evaporation. In many cases, as along the Great Rift Valley, the highlands are of volcanic origin, with weathered lava forming the basis for fertile, easily worked, and moisture-retentive red loams. In Kenya, Tanzania, and Uganda, cultivation has spread upward in such highlands with the introduction of temperate crops (in part furthered by European settlers), including especially the Andean, or Irish, potato, cruciferous vegetables of the genus *Brassica*, temperate species of peas and beans, and wheat and barley. The lower slopes are suited to Arabica coffee and the higher ones to tea and pyrethrum. This ascending wave of cultivation has pushed back the boundaries of the montane forests, which are now usually protected in forest reserves or national parks.

The presence of distinct agricultural zones at different elevations is most marked in Ethiopia, where the distinctive "false banana," or ensete, is grown at medium elevations in the forest belt of the south, Mediterranean fruits and vines are grown at higher elevations, and barley, wheat, and the indigenous cereal teff are grown in plowed fields on the high plateau.

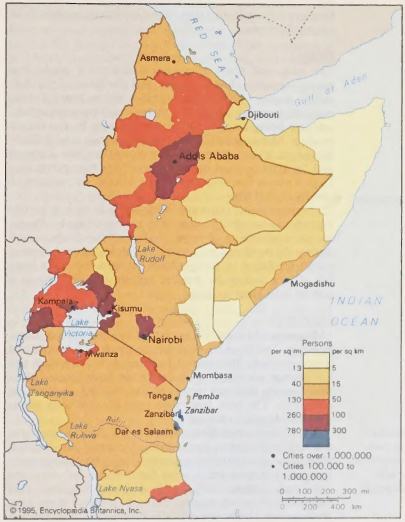
Cultivation in more arid regions. Regions with a lower annual rainfall or a pronounced dry season can only support the cultivation of less demanding, more drought-resistant crops such as sorghum, millet, and cassava. Commercial crops here include cotton and sisal. Where cattle can be kept, they improve the farming system, but, with huge areas infested by the tsetse fly, shifting agriculture is most prevalent. The agricultural problem in the drier regions is not so much a low annual rainfall as it is a long dry season, often lasting five to seven months. Another problem here is low soil fertility. Some fertility can be returned to the soil by slash-and-burn clearing and by a rotation of crops, but even this modest improvement is so quickly reduced that after a few years' cultivation the land must be allowed to return to bush fallow, preferably for 20 years or more. During this time the farmer works a sequence of clearings, which requires that he move his homestead or waste energy in long walks to his plots. With low population densities, this system provides a sustainable livelihood and a good return on the labour involved, but the scattered homesteads and their periodic abandonment make difficult the organization of modern marketing, transport and communication, and welfare services. On the other hand, greater population densities would lead to a clearing of the bush in which the tsetse flies breed, but the land so cleared would be unable to support the larger population.

Areas of typically marginal rainfall (*e.g.*, 20 to 30 inches per year) rely on a mixture of cultivation and livestock herding. In some years, which may be a majority, there is sufficient rainfall to bring a satisfactory harvest. Unfortunately, these years of agricultural plenty attract immigrants from overpopulated districts, who increase the scale of disaster when drought returns. Also, increased cultivation reduces plant cover and accelerates erosion, which further aggravates the situation. In this way, increased population intensifies the consequences of a normal climatic variation, leading in turn to an increased perception of drought and famine.

Recurring drought and famine

Livestock raising. Over large areas of eastern Africa, rainfall is inadequate for crop cultivation. This applies to the whole of Somalia and to some 70 percent of Kenya, which receive less than 20 inches in four years out of five. In areas such as these, the only feasible basis for land use is pastoralism. In the driest areas along the Red Sea coast, the whole of Somalia, and northeastern Kenya, the principal animal is the Arabian camel; elsewhere, cattle are dominant, usually in association with herds of sheep and goats and a few donkeys. These animals are multipurpose bases of livelihood, providing meat, milk, blood, hides, wool or hair, and transport. Since they are dependent upon grazing and browsing, they must be kept on the move to follow seasonal and other variations in rainfall, upon which the availability of vegetation and water depend. This nomadic way of life limits the accumulation of goods and chattels as well as the provision of such welfare services as schools and hospitals. The inoculation of livestock against disease and the construction of boreholes and reservoirs have enabled the people to increase the size of their herds, but this has led to overgrazing, with a consequent degradation of the rangeland and increased mortality in drought years. The problem here is that ownership is vested in the animals rather than the land, which is the primary resource. Therefore, it is in no one's interests to conserve grazing land, because it will only be used by someone else's herd. Also, the pastoralist is aware that experimental conservation could wipe out his herd and leave him helpless in this harsh environment. Individual or group ranches and systems of licensing have been set up to solve this quandary, but these conflict with cultural traditions and have not been very successful.

Irrigation. The irrigation of arid areas is limited by the amount of water that can be brought in from outside the



Population density of eastern Africa.

region, but not much of even this limited potential is utilized. For example, 70 percent of Kenya is cultivable only by irrigation, but only 3 percent receives more than 50 inches of rain, the minimal amount from which any considerable runoff can be expected. Only the Tana and the Athi–Galana river systems succeed in reaching the sea from the highlands, and irrigation schemes here are small. Developments in Somalia and in the Ogaden region of Ethiopia are effectively confined to the Jubba and Shebele rivers, which drain the eastern highlands and are used to grow bananas for export as well as cotton and food crops. In Ethiopia the waters of the Awash River are used as they descend from the highlands onto the floor of the Rift Valley; sugar and cotton are the major crops.

Irrigation in less arid zones

Irrigation is more promising in the less arid zones, where it can be used during the dry season or as a supplement to even out natural fluctuations in precipitation. This is done on sugar plantations in Uganda and on coffee estates in Kenya and Tanzania. Studies suggest that, if the floods of the rainy season on the Ruaha–Rufiji system of Tanzania were controlled, ample water would be available for irrigation in the dry season—as has already been demonstrated in small schemes in the Kilombero valley.

FORESTRY

The role of forests as a natural source of timber is confined by their small original extent and by deforestation. Forestry policies have been as concerned with the protection of watersheds as they have with production. Exports of natural hardwoods are very limited, although Uganda and Ethiopia can supply a modest domestic market. Local demands for softwoods are largely met by plantations of species of cypress (*Cupressus lusitanica* and *C. macrocarpa*) and pine (*Pinus radiata* and *P. patula*) derived from Central America. Black wattle (*Acacia mollissima*), introduced from Australia, is widely grown for firewood, and its spread has been greatly encouraged by being grown as a crop for tannin bark. The most widespread introduction from Australia, however, has been the eucalyptus, which, under eastern African conditions, grows very rapidly. Almost universally grown for firewood and poles, eucalyptus trees are a conspicuous part of the landscape, especially in upland Ethiopia.

FISHING

The lakes and rivers of eastern Africa are a productive natural resource. Lake Victoria and the lakes of the Rift valleys support fishing communities whose more distant markets, formerly supplied with sun-dried or smoked fish, are now being reached on a growing scale by the frozen product. Management of fish stocks has presented difficulties where lakes are bordered by more than one country, and the controversial introduction of the Nile perch into Lake Victoria has, since the 1980s, altered the balance of species in that body.

Inshore fisheries along the coast suffer from a generally narrow coastal shelf and a poor nutrient supply, except for some upwelling of deeper waters off the Somali coast during the northeast monsoon. The more remote oceanic waters are fished by foreign boats for tuna and other large fish. Game fishing for marlin, sailfish, and the like is a part of the tourist industry.

RESOURCES

Minerals. Mineral resources have so far proved disappointing. The lavas that blanket so much of Ethiopia and Kenya provide little except building stone, although the hot springs associated with volcanism have formed alkaline deposits that can be mined for soda ash and similar chemicals. The only large-scale extraction of soda ash is at Lake Magadi in Kenya.

Scarcity of commercially viable deposits

The ancient crystalline rocks of the African platform are rich in minerals that, having separated out through igneous or metamorphic processes into excellent geologic specimens, have a retail value for tourism, but there are few large, commercially viable deposits. Scattered gold finds have attracted mining, but only some of them have persisted for any length of time. Part of the most productive auriferous formation, exposed and worked in western Kenya and in Tanzania's West Lake region, lies beneath Lake Victoria. Lead and copper are among other minerals that have been mined in the past, but the most important continuing production has proved to be diamonds from one of the many Kimberlite pipes at Mwadui in Tanzania. A considerable deposit of iron ore has been proven to exist in southern Tanzania, but technical problems and difficulties of location and marketing have hindered development.

Also in southern Tanzania are deposits of coal. Exploration for oil and gas in the Red Sea region, the Tana River basin, and along the coast and offshore islands of the Indian Ocean have shown favourable indications.

Hydroelectricity. Although fossil fuels have to be imported, the majority of the region's commercial energy requirements have been met by hydroelectric power, which has considerable scope for expansion. Uganda's Owen Falls Dam provides power to both Uganda and Kenya, and it can be replicated at other sites along the Nile. Kenya has a major scheme on the upper Tana River and smaller plants and potential sites in the highlands, although the total surface flow is limited. Only a small portion of the Tanzanian potential is harnessed; the Ruaha–Rufiji system offers some good dam sites, but they would be expensive to exploit. In the high-rainfall area of the Ethiopian highlands, the great range of elevation has provided opportunities for power generation that have not fully been taken up. Only arid Somalia and Eritrea are not in a position to develop hydroelectric power.

COMMERCE AND INDUSTRY

Because the resource base for manufacturing industries is limited, the bulk of industrial exports are raw materials that are processed before shipment. Some industries based on domestic demand for such products as cement have taken enough advantage of large-scale production methods to export their products abroad. Manufacturing for local markets, over and above supplying food and beverages, takes the form of import substitution—that is, the manufacture, often from imported parts or raw materials, of goods that were once made abroad. Import-substitution industries are most successful in relatively large and affluent urban markets. In eastern Africa, this concentrates manufacturing in the capitals, which, being the largest cities and the centres of commercial functions as well as the preferred sites of international agencies, present the only areas of great demand that can sustain manufacturing activity.

Tourism

Tourism has seen different development among the East African countries (in the countries of the Horn, it is an insignificant sector of the economy). The industry is most important in Kenya, where receipts from foreign tourists are equivalent to income from a major export. Although tourism in Kenya is based on tropical beaches and on wildlife in national parks, these attractions also are found elsewhere, and contributory factors to Kenya's success are good international air connections, investment in hotels, roads, and other infrastructure, and political stability. Government policies in the other countries have been less supportive. (W.T.W.M.)

The people

EAST AFRICA

European knowledge of the peoples of the interior of East Africa began only in the second half of the 19th century, although knowledge of the coastal fringe had begun earlier, in the years after the Portuguese first made contact with Mombasa in 1498. Since the penetration of the interior most people, whether Europeans or East Africans, who have attempted to improve their comprehension of East African peoples have been struck and confused by the number of different named groups that make up the total population of the area. There are no reliable figures for the population of East Africa at the beginning of the 20th century, but, since there has been a very rapid increase since 1920, it is likely that the population in 1900 was less than 10 million. This population, however, was divided unequally among more than 160 distinct peoples—well

over this figure if a less conservative method of counting them is employed.

Identifying and classifying peoples. Although there is no complete agreement among ethnological specialists as to how a "people" or "ethnic group" is to be defined, there is substantial agreement among Africa specialists as to the vast majority of the identifiable groups in East Africa. For the purposes of this discussion, a people or ethnic group is a group of human beings who recognize their own identity and unity, have a name for themselves, and do not feel that they lose that identity in a larger grouping. Some groupings that now have the appearance of peoples, such as the Kalenjin of western Kenya, have come into being since 1960 by a conscious fusing together of older and smaller peoples. This series of fusions had not begun as early as 1900, although it is a safe speculation that most, if not all, of the peoples of that time also owed their existence in part to fusions of smaller groups at some earlier stage. For the purposes of this discussion, the year 1900 will mark the time by which East African peoples had retained practices long enough to be considered traditional and had not yet been disturbed by contact with Europeans.

Despite the fragmentation of the population into so many subdivisions, the different peoples of East Africa traditionally shared much of their cultures in common, thereby forming a smaller number of types, each type distinct in its characteristics. Africa scholars have attempted to identify and classify these types according to criteria of varying usefulness. The criteria discussed here are the following: descent, religion, language, habitat, subsistence, and political organization.

Criteria used to classify peoples

Descent. Categorizing peoples into groups according to their systems of descent and inheritance was much emphasized by writers between 1890 and 1950. Matrilineal descent—that is, formally reckoning family ties through the mother—occurred in the south (for example, among the Yao of Tanzania) and in a few pockets farther north, but most East African peoples formally reckoned descent patrilineally—that is, through the father. Bilateral descent as practiced in Europe was rare and possibly always of recent origin—that is, just before 1900.

Religion. By the late 19th century both Islām and Christianity were becoming widely known. But even before that time, most of the East African peoples took for granted a metaphysical model in which a supreme deity created and maintained the universe, and in which the spirits of dead ancestors watched over the prosperity and morals of each community and punished any offenders. In addition, the wild places were full of spirits, whose activities were unpredictable and often dangerous to humans. In order to cope with all of these mysterious powers, individuals, households, and communities consulted diviners and performed sacrifices of domestic animals. (Human sacrifices were rare.)

There were two basic variations from this model. Among the agricultural peoples living between Lakes Victoria, Albert, Edward, and Tanganyika in Uganda and northwestern Tanzania, a number of lesser gods received sacrifices along with the creator, the ancestors, and the spirits. But among the pastoral peoples of the northeast, including the Masai of southern Kenya and northern Tanzania, no attention was paid to ancestors, spirits, or gods, all devotion being directed to the creator alone.

All East African peoples were aware of the danger of witches, but, while some groups lived in terror of the next attack, others assumed that no witches lived near them.

Language. From about 1890 to 1960, East African ethnic groups were usually classified according to the affinities of the languages that they spoke. This was a very tidy method, but it gave rise to serious misapprehensions, since languages do not necessarily correlate closely with other features of culture. Indeed, language distribution, if anything, overemphasizes the diversity of East African peoples. For example, the Ganda of Uganda and the Kikuyu of Kenya speak very similar languages but are markedly distinct in traditional social organization, while the Masai resemble the Kikuyu closely in many traditional cultural details but speak an unrelated language.

The Kikuyu language belongs to the Bantu family, which covers the western, southern, and coastal areas of East Africa, while the Masai tongue belongs to the Nilotic family, which extends through the centre and north of the region. A third language group covering a large area is the Cushitic family, which includes Somali and Oromo and extends from the northeastern part of East Africa into the Horn of Africa. Apart from these, there are four or more other families, each represented by one or a few examples, but all of these four appear to have lost ground over time to the three major families.

Bantu, Nilotic, and Cushitic tongues

Habitat. A more useful way of grouping the East African peoples into types is according to their habitats, which can be summarized as follows: wet lowland, wet highland, semiarid, and arid. Wet lowland habitats are concentrated around Lake Victoria, and among the peoples found there about 1900 were the Ganda and Luo, both large in number. Wet highland habitats are less concentrated than are the lowland versions; they are strung out along the highlands of the western and eastern branches of the East African Rift System, and they occur also on a few large volcanic cones such as Kilimanjaro and Mount Kenya. The Kikuyu, living near Mount Kenya, exemplified the inhabitants of this type of habitat.

In East Africa wet country is fairly scarce, and most peoples have lived in semiarid country with a marked (but unreliable) wet season and a long dry season. The Kamba of Kenya and the Nyamwezi of Tanzania traditionally lived in different variants of the semiarid habitat, while in the arid country of northeastern Kenya lived the Somali and a few other groups. Much of this driest country was true desert, with no complete vegetation cover. The Masai, not very numerous but covering a large area, lived in a less austere version of Somali country.

Subsistence. Classification according to the four main types of habitat provides a more manageable number of peoples than the 160 named groups, but this method has the weakness of grouping together peoples that live in the same type of country and yet are marked by substantial differences in culture. Taking this complication into account, another way of classifying peoples is by their method of subsistence, that is, the physical basis of their survival. Again, four main types can be recognized among the peoples as they lived about 1900: hunters and gatherers, pastoralists, desultory cultivators, and intensive cultivators.

In 1900 only two peoples seem to have been hunting and gathering societies, divided into small bands among whom the men hunted larger wild animals while the women provided most of the food by gathering wild produce, most of it of vegetable origin. These two peoples were the Okiek of Kenya and the Hadza of Tanzania. They had no domestic animals except dogs and rarely, if ever, grew cultivated crop plants.

By contrast, pastoralists covered a much larger portion of East Africa, concentrating in the most arid areas of the north and northeast and extending through semiarid areas toward the centre of what is now Tanzania. The Masai, extending farther south than other pastoral peoples, kept goats, sheep, and cattle, with their value system revolving particularly around cattle. Most of the other pastoralists also placed an extreme value upon cattle, but, with the aridity of the country increasing toward the north and northeast, camels became more important to the survival of pastoral groups and, in the deserts of the Horn, replaced cattle as the high-prestige livestock among the Rendile and Somali. It is likely, but impossible to check, that the intense valuation of camels was derived from the Middle East, with cattle traditionally being the high-prestige livestock south of the Sahara.

Pastoralism

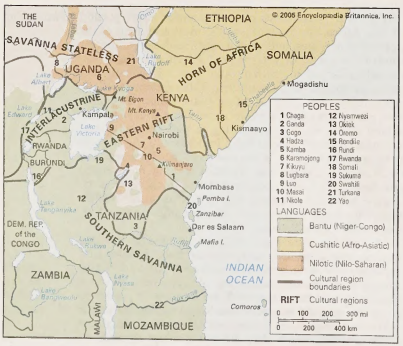
Although the term pastoralist can be used in a broad or a narrow sense, it is used here to mean peoples who depended for their survival on their herds and flocks. Among these groups, cultivation of crops was absent or ephemeral, and they relied little on hunting or gathering. However, they did rely very heavily on trading with their cultivating neighbours for an essential supply of grains and vegetables, which they did not produce themselves. In such trade, the high value of livestock and animal

products gave the pastoralists a strong bargaining position. However, they were also extremely vulnerable, because when drought or epizootics (epidemics among livestock) reduced their living capital they had no reserves on which to fall back. In such circumstances (apparently not infrequent), their main remedy was for the young men to run off somebody else's dwindling resources, and this is probably the reason for the pastoralists' reputation among their cultivating neighbours as ferocious raiders of herds and flocks. However, although the pastoral groups saw themselves as lords of creation, their distribution indicates that they lived only in areas where rain was too scarce or too unreliable to raise crops every year, if at all.

Scarce and unreliable rainfall were problems not only of the pastoralists; most of the cultivating peoples also lived with recurrent drought as a constant threat. An important buffer against the worst effects of drought was the buildup of herds and flocks, much as the pastoralists did. Then, in a dry year, the animals could be moved to water, whereas crops could not. Such was the high value that some cultivating peoples placed upon their livestock that they had much the same attitudes as pastoralists: they relied upon livestock as reserve wealth, identified themselves with their herds, and brought their livestock into their colour symbolism and songs. Examples of such peoples were the Kamba and Gogo of Kenya and Tanzania, respectively, and the Karamojong and Jie of Uganda. These peoples can conveniently be characterized as "would-be pastoralists." Their attitudes, however, did not impress the exclusive pastoralists—for instance, the Masai—who regarded their Kamba and Gogo neighbours as degraded by cultivation. Equally degraded in their eyes were those Masai who, having lost some or all of their livestock, found themselves forced to cultivate on the margins of the Masai plains, where rainfall was just adequate for an unreliable return. These Masai became, for the time at least, "would-be pastoralists."

Desultory cultivators

The desultory cultivators were peoples who lived in semi-arid country and put only a limited effort into working the soil, probably because they had a rough sense of the cost-benefit balance involved and realized that much of such effort would be lost to drought and, even in wet years, to pests such as insects and birds. There were even extra hazards associated with livestock. Over much of the northern half of East Africa, cattle-raising could be combined with cultivation. But over much of the southern half, in what is now Tanzania, the type of dry woodland often called miombo long has harboured tsetse flies, which carry infectious protozoans of the genus *Trypanosoma* that kill off cattle in these areas. Goats, sheep, and chickens could survive much better there, but in the southern half of East Africa it seems that, about 1900, most cattle were in the highlands, above the main tsetse areas.



Cultural and linguistic regions of East Africa.

Throughout East Africa, the highlands provided some of the best opportunities for intensive cultivation, because the rainfall there was relatively abundant and reliable. Similar opportunities occurred in lower areas of high rainfall, as around Lake Victoria, and on parts of the coast. An additional asset was fertile soil, since most East African, and indeed most tropical, soils are poor in plant nutrients. Soils formed locally over volcanic deposits, as in Kikuyu country, or over alluvial deposits, as in Luo country of western Kenya and north-central Tanzania, have a fertility that made possible dense populations, but only over relatively small areas. Intensive cultivation was the exception, not the rule, because the opportunities to make it work were exceptional. Where the opportunities occurred, however, cultivation repaid the hard work of tillage with hoes, making it worthwhile to hoe in great quantities of cattle dung and even to irrigate, as did the Chaga on Kilimanjaro.

Political organization. Contrasting the various modes of subsistence refines the classification of peoples based solely upon the dryness or wetness of their habitats. One further refinement, however, is necessary, and this is the contrast between those areas where the peoples were organized into states and those where they were not. The East African states were of various sizes, but all were characterized by hereditary heads of state (like European kings), by distinctions of social class, by formal administrative procedures, and, most vital of all, by procedures for collecting the taxes or tributes upon which each state's survival depended. Of the states, the largest in population were Buganda (the 19th-century kingdom of the Ganda people), Zanzibar (part of modern Tanzania), and Rwanda and Burundi.

Stateless peoples

By contrast with the states, in huge areas of East Africa the local peoples had no kings, no social classes, and no taxes. Everyone was a member of a local community but not of a larger administrative unit. The main social distinctions among these people were based on age, sex, and ability. Each local community's members conducted their own relations with other communities, and, if rival communities fought each other, the available fighting men might number from perhaps 15 to 50 on each side—whereas the kings of the largest states could put thousands of men in the field. Needless to say, local communities had difficulty resisting the might of kings. One example of such kingless and stateless peoples was the Masai.

Cultural regions. The clear and compact distribution of states and stateless societies provided a firm basis for the division of East Africa, in about 1900, into a limited number of distinct cultural regions. All the peoples of the south and west lived in states, while those of the north and northeast were stateless. None of the pastoral peoples had states, and neither did the few hunting and gathering peoples, so that the distribution of states and stateless societies largely reflected differences in the organization of the cultivating peoples. Among these, examples of both desultory and intensive cultivators were organized into states, while both types of cultivation were also found among the stateless peoples. For example, the Kikuyu had no states but were intensive cultivators, while the Chaga, who were equally intensive cultivators, did have states. Among peoples without intensive methods, the Gogo had no states—except in the west, where they adjoined the Nyamwezi, who had states and were equally unintensive in their cultivation methods.

In order to refine this scheme of cultural classification based upon the presence or absence of states, reference can be made to the custom of age-sets, and the distinctions that emerge thereby can be strengthened by further reference to the practices of circumcision and clitoridectomy. Age-sets were general throughout the northeast of East Africa, for example, among the Oromo and Masai. They were groups of males of roughly the same age who formally came of age during the same period and were for the rest of their lives ritually bound together, rather like adopted brothers, with all the other men in their set. In each area the sets formed a graded series, oldest men at the "top" until they retired and youngest men at the "bottom." Older always had authority over younger. Ini-

tiation into a set in most but not all areas was by way of circumcision, the cutting off of all or part of the foreskin of the penis. In most areas where circumcision was practiced, there also occurred "female circumcision," although there were no age-sets for women. In the female rite, the clitoris was cut out (clitoridectomy), and often the labia minora were also removed.

Five cultural regions of East Africa

The presence or absence of clitoridectomy, circumcision, age-sets, and states makes it possible to recognize five traditional cultural regions in East Africa, each of which possesses its own characteristics. These five are the Horn of Africa region, the Eastern Rift region, the Savanna Stateless region, the Interlacustrine region, and the Southern Savanna region.

The Horn of Africa. Although most of this region lies north of East Africa and is therefore discussed separately below, it does extend into the northeast part of Kenya. The Horn has closer cultural affinities with Arabia than with the four sub-Saharan regions of East Africa, owing to a long history of literacy, of large imperial states, and of Islām and Christianity—all of which are relatively new features in most of the other regions. In addition, the plow has long been an important aid to cultivation in the Horn, and the one-humped Arabian camel is an ancient domestic animal there. Most of the people of the Horn bear physical traits of the European geographic race, looking rather like dark-skinned Mediterraneans, while most people of sub-Saharan East Africa are of the African geographic race.

That part of the Horn within East Africa was inhabited, about 1900, by stateless pastoralists who kept in contact with nearby cultivating peoples in the Ethiopian highlands and in the river valleys of Somalia. Among these pastoralists, the Oromo kept cattle and camels, the Somali kept mostly camels, and both groups kept sheep and goats. The Somali were expanding at the expense of the Oromo, absorbing them culturally, so that increasingly more Oromo groups were losing their identity and becoming Somali. The Oromo had age-sets, but age-sets were maintained by only a few Somali groups in the south, where they doubtless had recent Oromo ancestors. The Somali circumcised because they were Muslims, but circumcision probably predates Islām in the Horn; certainly the non-Muslim Oromo practiced it long before 1900. Parallel to circumcision was clitoridectomy, there in a specialized form known as infibulation, in which, after the excising of the clitoris and labia minora, the opening to the girl's vagina was sewn up, leaving only a small aperture until she was married, when the vaginal opening was widened again.

The Eastern Rift. Although widespread in the Horn of Africa, infibulation did not spread to the four sub-Saharan regions. In the Eastern Rift region, however, clitoridectomy and circumcision were practiced, and male age-sets were found in all its areas except some border zones. Because age-sets, circumcision, and clitoridectomy were absent west and south of the Eastern Rift, it is tempting to speculate that these three cultural features spread southward into the region from the Horn. However, in the absence of written records before the arrival of Europeans, this theory is merely inferred from the distribution of the practices, and there is no positive evidence for it. Eastern Rift peoples had no states or kings, and they were well exemplified by the Kikuyu, the Kamba, and the Masai.

The Savanna Stateless. As its name implies, the Savanna Stateless region also lacked states, but it was set off from the Eastern Rift region by the absence of circumcision and clitoridectomy. (This region, it must be said, extended westward along the northern savanna of Africa into parts of West Africa, and there circumcision and clitoridectomy were found in places. However, in East Africa the contrast between the two regions was clear.) Age-sets did occur in the east of the Savanna Stateless region—for example, among the pastoralist Turkana and the "would-be pastoralist" Karamojong—but they were absent west of these peoples, so that the Lugbara and their neighbours in northwestern Uganda did not practice them.

The presence and absence of age-sets

The Interlacustrine and Southern Savanna. South of the savanna, the Interlacustrine peoples (those peoples living between the lakes now named Victoria, Albert, Edward, and Tanganyika) had states and kings, as did the ethnic groups of the Southern Savanna region, which covers most of western, southern, and coastal Tanzania. However, these two regions differed from each other in their traditions of kingship. Among the Interlacustrines, kings and kingdoms were linked to great ceremonial drums, more than one per state, which held the vitality of their state within them. This link was lacking in the Southern Savanna states, where a variety of locally important insignia marked off the kings from their subjects. (No African kings wore crowns.)

Another distinction between the two regions was that the Interlacustrine peoples, with few exceptions, were divided into castes of differing social prestige. In theory, but apparently not in practice, people of one caste were not allowed to marry members of another. Ranking of castes can be illustrated by the kingdom of Rwanda, where the Tutsi had higher prestige than the Hutu, and the Hutu were above the Twa. Among the Nkole peoples of western Uganda, the Hima were above the Iru. The Ganda, however, did not have castes. Southern Savanna peoples, all of them casteless, included the Nyamwezi and the Yao.

In these two regions (at least in those parts within East Africa) there was traditionally no circumcision or clitoridectomy. However, the Southern Savanna region extended westward across the continent to the Atlantic Ocean, and in the west, circumcision was practiced over a large area. Another complication, apparent by 1900, was that, as increasingly more men became Muslims, they were circumcised. The Swahili peoples of coastal East Africa had long been Muslim and had long been in contact with the Middle East, but with their penetration of the interior during the 19th century, Islām also spread and brought with it the new practice of circumcision. (J.D.K.)

THE HORN OF AFRICA

The Horn of Africa, an extension of land between the Indian Ocean and the Gulf of Aden, is occupied by Ethiopia, Eritrea, Somalia, and Djibouti, whose cultures have been linked throughout their long history.

Principal ethnic groups. Ethiopia has a history of independent sovereignty extending at least 2,000 years. Discounting the brief Italian intrusion (1935–41), it was the only traditional empire to survive the colonial partition intact. Until the mid-1970s Ethiopia was ruled by leaders drawn almost exclusively from the two dominant ethnic groups, the Semitic-speaking Tigray and Amhara. Ethiopia is thus essentially the political expression of the cultural nationalism of these two closely related peoples, who derive from a fusion of local Cushitic-speaking stock with South Arabian immigrants who settled along the Red Sea coast in the 1st millennium BC.

The Tigray and Amhara

The Tigray occupy the northern part of the Ethiopian Plateau. This highland, which straddles the Ethiopian-Eritrean border, contains the ancient capitals of the empire: Aksum, Gonder, and Lalibela. For more than half a millennium, however, power lay mainly with the Amhara, who live in the southern part of the plateau in the administrative regions of Gonder, Gojam, and Shewa—the seat of government since 1889. The historic rivalry between these two groups has been reflected most recently in rebellions against "Amhara-dominated" Ethiopia by nationalists in the region of Tigray and by secessionists in Eritrea.

Prone to recurrent drought and famine, the heavily overused Ethiopian Plateau and Eritrean highlands produce the indigenous cereal teff (*Eragrostis abyssinica*), as well as wheat, barley, and, in the drier regions especially, corn (maize) and millet. Cattle and other stock are raised, the land being tilled by ox-drawn plow. In the lower and hotter southern regions of Ethiopia, the false banana, known locally as ensete (*Ensete ventricosum*), is the main crop, and coffee—which is indigenous to Ethiopia and may be named for a local tribe, the Kafa—is produced as a cash crop. There, for the most part, the hand hoe replaces the ox-drawn plow in cultivation. This ensete-growing region is the home of the Gurage (a Semitic-speaking but partly Cushitic people), of the Cushitic Sidamo-speaking peoples, and, to the southwest, of a few other, more distantly connected peoples, whose linguistic affiliation remains a matter of controversy.



Peoples and language areas of the Horn of Africa.

By far the most important group among the Cushitic people is the Oromo, who form the largest ethnic unit in northeastern Africa. They occupy most of the southern provinces of Ethiopia, with the related Somali on their eastern and southern flanks, and seem destined to play an increasingly significant role in the political development of Ethiopia. Their traditional pastoral nomadism is best preserved today among the Boran Oromo, who live in the hot, dry lowlands of southern Ethiopia and northern Kenya. It is from this region that the many different subdivisions of the expanding Oromo nation began their invasion of Ethiopia in the 16th century, sweeping relentlessly in wave after wave into the Christian highlands, where the embattled Amhara and Tigray were unable to check their advance. Those Oromo who moved into the rich central highlands, such as the Macha and Tulama, abandoned their nomadic economy and became sedentary cultivators; others, such as the Arusi and Gudji, in the less-favoured areas sought to combine both modes of livelihood. Many became Christians, adopting Amhara culture; others embraced Islām, although this was not necessarily incompatible with adopting other aspects of Amhara culture.

The inability of the Christians to withstand the Oromo invaders (over whom, however, they later reestablished their ascendancy and whom they finally incorporated as subjects in the empire) shows the extent to which Ethiopian fortunes had been reduced by the ceaseless wars of the period with the surrounding Muslim Sidamo and other largely Cushitic sultanates.

Somali

The Cushitic Somali became fervent Muslims and put increasing pressure on the southeastern flank of the Oromo, thus helping to sustain the latter's continued thrust into Ethiopia. After a long period of turmoil, adjustment, and the reemergence of Amhara ascendancy, under the forceful rule (1889–1913) of Menilek II, Ethiopia assumed its modern shape. Despite their crushing defeat at the Battle of Adowa (Adwa) by the Ethiopians in 1896, the Italians were allowed to retain the largely Muslim and partly Cushitic territory of Eritrea, which was joined to Ethiopia after World War II but became an independent state in 1993.

Notwithstanding their strong sense of cultural nationalism, the Somali had not previously formed a single political unit, and they were partitioned among the Ethiopians, Italians, British, and French. The French also acquired the closely related Cushitic-speaking Afar to form their tiny colony around the port of Djibouti. In 1960 the British and Italian parts of the Somali nation became independent and joined to form Somalia, leaving about a quarter of the total Somali population in the neighbouring areas of eastern Ethiopia, northern Kenya, and the minuscule Republic of Djibouti, which the Somali share with the Afar. Formerly known as French Somaliland, this arid land contains mainly nomads or urban workers.

Traditional ethnic and religious rivalries are thus perpetuated today by the coexistence of a state based on Somali identity (but not including the whole nation) and the ancient state of Ethiopia, which, like most of its more recently formed African neighbours, includes within its borders a variety of different peoples and tribes, about one-third of whom are Muslim and many also Cushitic.

Ethos. The worldviews of the Christian and Muslim peoples of the region are uncannily alike, and, even in particular aspects of belief and ritual, they reveal many striking resonances. Both believe strongly in a morally significant afterlife and are thus capable of accepting present misfortune and illness with fatalistic resignation. Both are also equally adept at seeking alternative mystical explanations of distress and of resorting to supernatural agencies to seek redress. Otherworldly fatalism is thus tempered by a healthy pragmatic concern for present well-being. Those Oromo and Sidamo who retain their traditional cosmologies (and few have remained entirely unaffected by Islām or Christianity) are more firmly anchored in the present and entertain few if any hopes of eternal bliss or fears of eternal damnation.

The most obvious contrasts in cultural outlook and ethos follow the division between cultivator and nomad. The Amhara and Tigray farmers of the highlands are sturdy, canny peasants whose ready deference to their social superiors conceals hostility and suspicion. Their strongly individualistic and shrewdly calculating attitudes suit their hierarchical but far from closed status system and kinship organization.

Specialist crafts, such as weaving, leatherworking, and ironworking, are traditionally despised, and their practitioners are associated with the evil eye. Other artisan work, unskilled manual labour, and even trade also are considered degrading. Specialist minority groups—such as the Dorse weavers or the Cushitic Beta Israel (the Falasha, or "Black Jews"), who traditionally do a considerable amount of ironwork and pottery—are thus able to establish ethnic monopolies. Similarly, the traditionally disparaged Gurage have acquired a leading role as manual workers in Addis Ababa, and, until quite recently, trade has tended to be monopolized by Muslims.

The ethos of the nomads reflects their egalitarian social structure. Their assertiveness contrasts strongly with the deferential respect of the Amhara peasantry. Pragmatic individualism is tempered by the wider demands of the kinship group. Intense and violent competition rages over access to the sparse resources of the environment—grass and water—on which life depends. Enmities and alliances tend to be ephemeral and shifting, the definition of friend and foe constantly changing. In this, the much-divided Somali, whose constituent factions (now armed with automatic rifles) are regularly embroiled in savage feuds, contrast with the Oromo, whose component groups place a high value on internal peace and count the killing of one of their own fellows as a sin. Yet all these warrior nomads—whether Somali, Oromo, or Afar—hold a similar heroic view of life, in which prowess in battle and raiding is the essential manly virtue. The weak and vulnerable receive a certain condescending compassion that is associated with the idea that they may possess special (compensating) mystical powers. In a world in which essentially might is right, however, true prestige is accorded only to those who are manifestly strong and successful. (The Amhara military aristocracy holds similar values.)

Warlike nomads

Kinship, descent, and age-sets. Groups based upon descent from a common ancestor play some role in the social organization of all the peoples mentioned. Generally, it

is patrilineal descent (traced through male ancestors on the father's side of the family) that is of most significance in the inheritance of property and status and in group formation. It is weakest among the Christian Amharas, who regard kinship links traced through either women or men as equally binding. Traditionally, this bilateral kinship system was employed to build up loosely defined clusters of kin each associated with a particular parish or part of a parish. The church-centred local community formed the primary unit in the Amhara administrative system, consisting of a series of scattered hamlets rather than a clearly defined village, and rights to land rested within exogamous kin groups. After the rise of the socialist regime in 1974, land reforms abolished these traditional rights. Control of the land was allocated to local peasant associations (called *kebelle*s), which, in the northern Amhara regions, were often chaired by priests.

Among the other mainly Cushitic-speaking peoples (including the Semitic-speaking Gurage), descent is more strictly patrilineal, and conventional kinship groups such as patrilineages and clans occur at various orders of social grouping. Where sedentary cultivation is practiced, as among the central and northern Oromo, the Gurage, and the Sidamo, there is a tendency for particular clans, or their constituent lineages, to be associated with particular localities. Communities may contain the members of several separate descent groups, but one of these is identified with a given locality.

Clans and age-sets

The extent of clan and lineage development and ramification is largely a function of the size of the tribal units involved. The Oromo, for instance, comprise at least eight major tribes, ranging from the most traditional and largely pastoral Gudji and Boran in the south to the strongly Amhara-influenced, cultivating Tulama and Macha around Addis Ababa. In the large units, kinship is only an ancillary principle of association, supplementing more important ties based on shared membership of a generation class, or age-set. According to the so-called *gada* system, all the Oromo male children of the same generation formed an indissoluble fraternity; and, as members of this fraternity, they moved through the various stages of life and positions open to them. The age-sets held in turn the statuses of bachelor-warriors and later of married elders, each set occupying a given grade for a period of approximately eight years. Each generation eventually supplied the lawgivers and leaders of a given Oromo tribe and then retired to make way for its successors.

This form of government was essentially democratic and republican since there was a constant rotation of office-holders and each age-set elected its own leaders who, when the time came, would rule the tribe as a whole. The traditional political system was altered radically, however, as the Oromo expanded northward and many groups changed their location and economy. Political changes also were influenced by the larger Amhara centralized system and by the adoption of either Islām or Christianity, which were often invoked to legitimize new structures of authority.

The *gada* organization provided the strong sense of cohesion that even large nomadic groups such as the Boran were able to achieve. Age-set organization and descent are also significant principles in the local organization of the Sidamo people. But it is among the Afar and certainly the Somali that descent becomes the most significant principle of social and political allegiance. Ties based on common residence in a given area count for less among the northern pastoral Somali than is true of any other group in northeast Africa, and patrilineal kinship is nowhere more important or more heavily utilized in group formation. Patrilineal descent, given specific range and content by contractual treaties, provides the key to the traditional Somali political and legal system. Such treaties result in the formation of distinct politico-legal units, containing a few thousand male kin (and their dependents) who have agreed to meet all liabilities in concert. On this basis, if the parties involved in certain deaths and injuries wish to avoid feuding, they can collectively adjudicate their differences through a system of mutual compensations that aim at a kind of balance of payments. So while lineages of the nomads are not identified with any locality, whenever the need arises scattered kinsmen rally together to attack their foes or defend their interests.

Settlement and livelihood. Village settlements are not found among the Amhara or among those who have been strongly influenced by their culture and way of life. The basic settlement unit in the rural areas is the nuclear or extended family (that is, the husband-and-wife family or its extension to kinfolk, respectively) living in a few circular mud-and-wattle houses. Formerly, the typical parish contained a core of some 20 to 30 household heads descended from a common ancestor associated with the parish and holding landrights in it. This traditional pattern of largely autonomous settlements was radically changed by the socialist government. After 1974, the chairmen of local *kebelle*s allocated land to members of the community, and they also selected people for military service, forced labour on state farms, or resettlement elsewhere. To break down the isolation of the peasants and bring them under the control of the central government, a villagization scheme was introduced in 1985. By the end of 1987, eight million people had been regrouped into separate village settlements, each containing up to 500 households. Larger settlements, with thousands of residents from different parts of Ethiopia, were created in an agricultural resettlement project intended to provide arable land and relief from famine.

The Amhara pattern of dispersed settlement is replicated fairly closely among the ensete-growing Sidamo as well as among most of the settled, cultivating Oromo. The Cushitic-speaking Konso with their clearly defined and traditionally fortified villages and the southern Somali with similarly distinct settlements are both unusual. Southern Somali villages are based upon communal water ponds to which access is strictly controlled by membership in the local community and by participation in its blood-compensation arrangements. Among all these groups, land is not traditionally transferable except to other members of the group or to those who join it as clients. Its use entails social and politico-legal obligations, and, when it is abandoned, it reverts traditionally to the local community.

Among the Cushitic pastoralists, settlement patterns are naturally more fluid and flexible. The Boran, who are less mobile and far-ranging than the Somali in their movements, have generally separate grazing encampments for each type of livestock. The Boran rank their animals in the following order: lactating cows, dry cows, lactating camels, dry camels, and sheep and goats. The most senior brother in an extended family herds the first category of stock in the grazing regions best suited to it, and other brothers look after the other animals in descending order of seniority and prestige.

Seasonal movement of nomadic herds

The pastoral Somali nomads, on the other hand, have basically only two types of herding unit. The first consists of the camels, which in the dry seasons can go without water for up to 20 days or more. These are in the charge of young unmarried men and, in the dry seasons particularly, seek good grazing, sometimes hundreds of miles from the wells to which other stock—sheep and goats, cattle and milk camels—are compelled to cling closely. These latter are herded by the family unit, consisting of a man and his wife or wives and their unmarried daughters and young sons (boys over the age of seven are out with the grazing camels learning the techniques of camel management). The family moves from pasture to pasture carrying its readily transportable tents and other equipment on burden camels, which are not, however, ridden. Pasture is not owned, and rights to wells are asserted with an intensity that increases in direct proportion to the scarcity of water and the energy expended in utilizing it. In the wet seasons, when lush pasture regions become centres of intensive grazing and settlement, the two herding units move closer together, and social life becomes more expansive and relaxed. With milk and meat in abundance, this is the time for feasts and collective rituals, the season of marriage negotiation and weddings.

Except among the Christians, all these peoples practice polygyny in one form or another, each wife and her children usually forming a separate domestic and economic unit. Marriage payments, largely in livestock among the

pastoralists, are probably highest in the case of the pastoral Somali. Among these Somali, however, a bride also brings a considerable dowry in the form of a flock of sheep and goats to provide milk for the children whom she is expected to bear her husband.

Finally, a striking feature of Ethiopian economics is the complex and overlapping system of regional markets held in a given centre on a particular day of the week, like the pattern that still obtains in parts of Europe. Saturday is the most popular day. Although such large centres as Gonder or Addis Ababa have daily markets, the weekly market is usually much bigger and has its own distinctive festive atmosphere.

Religion. The Cushitic tradition, traces of which occur even among people as long and deeply Islāmized as the Somali, seems best preserved today (in at least one of its original or early forms) among the Boran. Here Waqa, the god of sky and earth and the creator and sustainer of life, is worshiped in prayer and sacrifice as the guardian of social morality and as the source of all things, good and bad. Waqa's special agents on earth are the sacred dynasties, or lineages, of priests (*kallus*), who still live among the Boran and to whom all the Oromo in ancient times used to send emissaries on pilgrimage. The pilgrims came to receive the blessing of the *kallu* priests, or "anointing fathers," who thus made sacred the whole traditional Oromo social system. Today, among the Macha and other Oromo who now live in the Ethiopian highlands, these traditional national priests have been replaced by new spirit-possessed charismatic leaders (also called *kallus*) who express the ethnic identity of their tribesmen in the context of the Christian Amhara-dominated Ethiopian state.

The Ethiopian Orthodox Church

The Ethiopian Orthodox Church maintains the monophysitic doctrine that Christ has a single nature into which his divine and human sides are assumed. It follows the Alexandrian rite, its liturgy being celebrated and recorded in the ancient Ethiopian language Ge'ez (from which Amharic is derived). It is led by the *abuna*, an Ethiopian priest who before 1948 was appointed by the Coptic patriarch in Alexandria but since has been appointed locally. The *abuna* presides over a church that, with its numerous places of worship, its richly endowed monasteries, and its ample priesthood, held perhaps one-fifth of Ethiopia's arable land before the revolution. Its all-pervasive character is such that one of every five male Christians is estimated to be in orders, and priests are expected to be married, only monks and nuns being celibate. Unordained, but not necessarily unlearned, ritual experts called *debtaras* play a crucial role in dispensing mystically efficacious cures. There is a vast hierarchy of saints and angels, chief among them being the Virgin Mary, St. Gabriel, St. Michael, St. George (patron saint of Ethiopia), and the local saints Tekle Haimanot and Gabra Manfas Keddus. On the darker side, a complementary host of demons and evil spirits, many connected with a form of witchcraft (*buda*) and able to possess people and cause illness and even death, are widely feared. Though officially discouraged under the socialist regime, the protective cult of the saints is still vast and all-embracing. Every Christian has a special relationship with several saints, and the observance of saints' fasts and feasts bulks large in the Christian calendar. Spirits of every origin and provenance are accepted within the Christian cosmology, where they are naturalized in the continuous process of cultural exchange between the dominant and subject peoples. Thus, for instance, the Oromo fertility spirit Atete is readily assimilated to the Virgin Mary, and vice versa.

Exactly the same pattern is repeated on the Muslim side, where the cult of saints is equally well developed and the Islāmic cosmology coincides to a remarkable degree with that of the Amhara Christians. Thus, among the Somali, who posthumously canonize their own lineage ancestors, saints are petitioned to remedy every distress and anxiety and are venerated as essential mediators between man and the Prophet Muḥammad and God. Again, the process of Islāmization closely parallels that of Christianization: in the case of the Arusi Oromo, for instance, the Prophet himself and numerous other Muslim saints are assimilated to traditional spirits and ultimately to Waqa.

If the great traditions of Christianity and Islām generously open their arms to assimilate the many local cultures of the region, many elements from the local cultures in turn find their way, by the back door as it were, into the worldview of the two major religions. This is very clearly seen in the mystery spirit-possession cults, which attract women and certain underprivileged categories of men particularly and which have a strongly "underground" character. Christians are especially susceptible to harassment by Muslim and pagan spirits, just as Muslims are equally open to attack by Christian demons or pagan spirits. Indeed, one of the salient features of religion in this part of Africa is the emphasis placed on spirit possession and the evil eye as explanations of misfortune that elsewhere would be ascribed to witchcraft or sorcery. The other striking feature deserving notice is the stress placed on food taboos and other commensal restrictions that are applied with particular stringency to maintain social distance from hunting and low-status craft groups. The despised and rejected community, on the other hand, may entertain similar attitudes toward those who consider themselves their superiors. Thus, just as other Ethiopians traditionally look down on the religion of the Jewish Falasha and on the ironwork and pottery trades that they follow, the Jewish Falasha themselves seek to remain apart in order to preserve their own ritual purity. Those who have close contact with, or live among, non-Jews are treated with disdain.

Fear of spirits and demons

The presence for centuries of three world religions in the region, each with its own literate tradition, has preserved a more obvious and tangible literary heritage than in most other parts of Africa south of the Sahara, but the bulk of this literature remains largely untranslated and unknown to the outside world. More accessible are the splendid monuments of the Christian past: the celebrated rock-hewn churches of Lalibela, which were largely constructed during the reign of the 12th-century emperor after whom the city is named; the glorious castles and churches of Gonder with their famous, magnificently decorated ceilings; and, beyond this tradition, the older Aksumite antiquities. Islām, too, has left comparable memorials in some of the earliest mosques and tombs in such ancient cities as Harer, Seylac, and Mogadishu but, for religious reasons, has left no illustrations comparable to those adorning Ethiopian manuscripts and church walls. Outside these literate traditions, there exists a less well explored heritage of oral literature, in both prose and poetry. (I.M.L.)

History

EAST AFRICA

The coast until 1856. The earliest written accounts of the East African coast occur in the *Periplus Maris Erythraei*—apparently written by a Greek merchant living in Egypt in the second half of the 1st century AD—and in Ptolemy's *Guide to Geography*, the East African section of which, in its extant form, probably represents a compilation of geographic knowledge available at Byzantium in about 400. The *Periplus* describes in some detail the shore of what was to become northern Somalia. Ships sailed from there to western India to bring back cotton cloth, grain, oil, sugar, and ghee, while others moved down the Red Sea to the East African coast bringing cloaks, tunics, copper, and tin. Aromatic gums, tortoiseshell, ivory, and slaves were traded in return.

Early trade with Arabia

Azania. Because of offshore islands, better landing places, and wetter climate, Arab traders from about 700 seem to have preferred the East African coast to the south of modern Somalia. They sailed there with the northeast monsoon, returning home in the summer with the southwest. They dubbed the part of the coast to which they sailed Azania, or the Land of Zanj—by which they meant the land of the blacks and by which they knew it until the 10th century. South of Sarapion, Nikon, the Pyralaae Islands, and the island of Diorux (about whose precise location only speculation seems possible), the chief town was Rhapta, which may lie buried in the Rufiji delta of present-day Tanzania. Here the situation differed somewhat from that in the north, and, though tortoiseshell and rhinoceros horn were exported from there—as were

quantities of ivory and coconut oil—no mention is made of slaves. Rhapta's main imports were metal weapons and iron tools—suggesting that iron smelting was not yet known. Mafia Island, which lies out to sea here, could perhaps be Menouthias, the only island named in both the *Periplus* and the *Guide*, although this could also be either Pemba or Zanzibar (perhaps there has been a conflation of all three in the one name).

There is little information concerning the period until the 8th century. Greek and Roman coins have been found, and there are some accounts of overseas migrations to the coast. No settlements from this period have been found.

A new period opened, it seems, in the 9th century. The first identifiable building sites are dated from this time, and, according to Arab geographers, the East African coast was then generally thought of as being divided into four: (1) Berber, which ran down the Somalian coast to the Shabeelle River, (2) Zanj proper, (3) the land of Sofala in present-day Mozambique, whence gold was beginning to be shipped by about the 10th century, and (4) a vaguely described land of Waq waq, beyond. The only island that is mentioned is Qanbalu, which appears to have been what is now Tanzania's Pemba Island. Though there is some suggestion that in the 10th century the Muslims had not yet begun to move farther south than Somalia, on Qanbalu they soon became rulers of a pagan population, whose language they adopted. Moreover, at Zanzibar an extant Kūfic inscription (the only one) recording the construction of a mosque by Sheikh as-Sayyid Abū 'Imrān Mūsā ibn al-Ḥasan ibn Muḥammad in 1107 confirms that by this time substantial Muslim settlements had been established.

The main coastal settlements were situated on islands, largely, no doubt, because of the greater security these provided against attacks from the mainland; and their populations seem mostly to have been made up from migrants from the Persian Gulf—some from the great port of Sīrāf, others from near Bahrain—though conceivably some too came from Daybul, at the mouth of the Indus River, in northwestern India. They exported ivory (some of it went as far as China) and also tortoiseshell, ambergris, and leopard skins. Such trade goods as they obtained from the interior were apparently bought by barter at the coast.

Kilwa

Ruins at Kilwa, on the southern Tanzanian coast, probably date from the 9th or perhaps from the 8th century. They have revealed an extensive pre-Muslim settlement standing on the edge of what was the finest harbour on the coast. Though there is little evidence to suggest that its inhabitants had any buildings to begin with, wattle-and-daub dwellings appeared in due course, and by the 10th century short lengths of coral masonry wall were being built. The inhabitants, whose main local currency was cowrie shells, traded with the peoples of the Persian Gulf and, by the early 11th century, had first come under Muslim influence. By 1300, like the inhabitants of neighbouring Mafia, they were living in Muslim towns, the rulers of which were Shi'ites. Although no houses were being built of coral, stone mosques were being constructed.

External trade was increasing: glass beads were being imported from India, and porcelains, transshipped either in India or in the Persian Gulf, were arriving from China. There appears also to have been a rather extensive trade with the island of Madagascar.

The most important site of this period yet to have been found is at Manda, near Lamu, on the Kenyan coast. Apparently established in the 9th century, it is distinguished for its seawalls of coral blocks, each of which weighs up to a ton. Though the majority of its houses were of wattle and daub, there were also some of stone. Trade, which seems to have been by barter, was considerable, with the main export probably of ivory. Manda had close trading connections with the Persian Gulf—with Sīrāf in particular. It imported large quantities of Islāmic pottery and, in the 9th and 10th centuries, Chinese porcelain. There is evidence of a considerable iron-smelting industry at Manda and of a lesser one at Kilwa.

The Shirazi migration. For much of the 13th century the most important coastal town was Mogadishu, a mercantile city on the Somalian coast to which new migrants came from the Persian Gulf and southern Arabia. Of these, the most important were called Shirazi, who, in the second half of the 12th century, had migrated southward to the Lamu islands, to Pemba, to Mafia, to the Comoro Islands, and to Kilwa, where by the end of the 12th century they had established a dynasty. Whether they were actually Persian in origin is somewhat doubtful. Though much troubled by wars, by the latter part of the 13th century they had made Kilwa second in importance only to Mogadishu. When the Kilwa throne was seized by Abū al-Mawāhib, major new developments ensued. Kilwa captured Mogadishu's erstwhile monopoly of the gold trade with Sofala and exchanged cloth—much of it made at Kilwa—and glass beads for gold; and with the great wealth that resulted new pottery styles were developed, a marked increase in the import of Chinese porcelain occurred, and stone houses, which had hitherto been rare, became common. The great palace of Husuni Kubwa, with well over 100 rooms, was built at this time and had the distinction of being the largest single building in all sub-Saharan Africa. Husuni Ndogo, with its massive enclosure walls, was probably built at this time, too, as were the extensions to the great mosque at Kilwa. The architectural inspiration of these buildings was Arab, their craftsmanship was of a high standard, and the grammar of their inscriptions was impeccable. Kilwa declined in the late 14th century and revived in the first half of the 15th, but then—partly because of internal dynastic conflict but also partly because of diminishing profits from the gold trade—it declined again thereafter.

Prosperity in the 15th century

Elsewhere, especially on the Kenyan coastline, the first half of the 15th century seems to have been a period of much prosperity. Whether at Gede (south of Malindi) or at Songo Mnara (south of Kilwa), architectural styles were relatively uniform. Single-story stone houses, mostly of coral, were common. Each coastal settlement had a stone mosque, which, typically, centred upon a roofed rectangular hall divided by masonry pillars. Chinese imports arrived in ever larger quantities, and there are signs that eating bowls were beginning to come into more common use. Mombasa became a very substantial town, as did Pate, in the Lamu islands. The ruling classes of these towns were Muslims of mixed Arab and African descent who were mostly involved in trade; beneath them were African labourers who were often slaves and a transient Arab population. The impetus in this society was Islāmic rather than African. It was bound by sea to the distant Islāmic world, whence immigrants still arrived to settle on the East African coast, to intermarry with local people, and to adopt the Swahili language. The impact of these settlements was limited, while their influence upon the East African interior was nonexistent.

During the 15th century, Shirazi families continued to rule in Malindi, Mombasa, and Kilwa and at many lesser places along the coast. They also dominated Zanzibar and Pemba. The Nabahani, who were of Omani origin, ruled at Pate and were well-represented in Pemba as well. Coastal society derived a certain unity by its participation in a single trading network, by a common adherence to Islām, and by the ties of blood and marriage among its leading families. Politically, however, its city-states were largely independent, acknowledging no foreign control, and their limited resources confined their political activities to East Africa and to a variety of local rivalries—Zanzibar and Pemba, for example, appear frequently to have been divided between several local rulers. Mombasa occupied the premier position on this part of the coast, although its control over the area immediately to the north was disputed by its main rival, Malindi. Close connections seem to have existed between Mombasa and a number of places to the south. Its Shirazi rulers were able to mobilize military support from some of the inland peoples, and as a result of the place it had won in the trade of the northwestern Indian Ocean they had turned Mombasa into a prosperous town. Its population of about 10,000 compared with only 4,000 at Kilwa.

The Portuguese invasion. This was the situation on the East African coast when Portuguese ships under Vasco da Gama arrived in 1498. The manifestly superior military and naval technology of the Portuguese and the greater

unity of their command enabled them, in the years that lay ahead, to mount assaults upon the ill-defended city-states. As early as 1502 the sheikh at Kilwa was obliged to agree to a tribute to the Portuguese, as the ruler of Zanzibar was later. Shortly afterward the Portuguese sacked both Kilwa and Mombasa and forced Lamu and Pate to submit. Within eight years of their arrival they had managed to dominate the coast and the trade routes that led from there to India.

Ejection of the Shirazi

The Portuguese became skilled at playing one small state against another, but their global enterprise was such that they did not immediately impose direct rule. This changed toward the end of the 16th century, however, when Turkish expeditions descending the northern coast with promises of assistance against the Portuguese encouraged the coast north of Pemba to revolt. This prompted the dispatch of Portuguese fleets from Goa, one of which, in 1589, sacked Mombasa and placed that city much more firmly under Portuguese control. This was helped by the death of Mombasa's last Shirazi ruler, Shah ibn Mishhan, who, in leaving no clear successor, gave the Portuguese the opportunity to install Sheikh Aḥmad of Malindi in his place. In 1593, with an architect from Italy in charge and with masons from India to assist them, the Portuguese set about building their great Fort Jesus at Mombasa. In the following year it was occupied by a garrison of 100 men.

With Mombasa's downfall, the major hindrance to Portuguese power on the East African coast was overthrown. They installed garrisons elsewhere than at Mombasa and brought about the downfall of a number of Shirazi dynasties, and, although they did not exercise day-to-day control over local rulers, they did make them dependent on them for their position. (Local rulers were in particular required to pay regular tribute to the Portuguese king on pain of dethronement and even of death.)

Portugal's chief interests were not imperial but economic. With Mombasa in their grip, they controlled the commercial system of the western Indian Ocean. Customs houses were opened at Mombasa and Pate, and ironware, weapons, beads, jewelry, cotton, and silks were imported. The main exports were ivory, gold, ambergris, and coral. There was a flourishing local trade in timber, pitch, rice, and cereals but few signs of any considerable traffic in slaves. Individual Portuguese traders often developed excellent relations with Swahilis in the coastal cities.

Though the Portuguese managed to ride out local rebellions into the 17th century, their authority over a much wider area was undermined by the rise of new powers on the Persian Gulf. Portugal lost Hormuz to the Persians in 1622 and Muscat to the imam of Oman in 1650. Two years later the Omanis launched their first major intervention into East Africa's affairs when in response to a Mombasan appeal the imam sent ships to Pate and Zanzibar and killed their Portuguese inhabitants. As a consequence, Pate became the centre of East Africa's resistance to Portuguese rule. The Portuguese responded in an equally bloody manner, but eventually, in 1696, in alliance with Pate, the imam of Oman sailed to East Africa with a fleet of more than 3,000 men to lay siege to Mombasa. Although Fort Jesus was reinforced, the great Portuguese stronghold finally fell to Sayf ibn Sulṭān in December 1698. A few years later Zanzibar, the last of Portugal's allies in Eastern Africa, also fell to the imam.

The Omani ascendancy. There ensued, after the Omani victory, a century during which, despite a succession of Omani incursions, the East African coast remained very largely free from the dominance of any outside power. Oman itself suffered an invasion by the Persians and was long distracted by civil conflict. Its originally successful Ya'rubid dynasty lost prestige as a consequence, fell from power, and was then superseded by the Āl Bū Sa'īdīs, who very soon found themselves preoccupied by conflicts at home. Moves against them also originated along the East African coast.

The Āl Bū Sa'īdi dynasty

In 1727 Pate joined with the Portuguese to expel the Omanis, especially from Mombasa, where in 1728–29 Portuguese authority was momentarily restored. But the Mombasans wanted as little to be controlled by Portugal as by Muscat and soon evicted the Portuguese once again. Thereafter, Kilwa, Zanzibar, Lamu, and Pate largely kept themselves free from both Omani and Portuguese control. Distracted though it was by protracted internecine quarrels, Pate was preeminent in the Lamu archipelago and, like all the other coastal towns, was ambitious to preserve its independence.

Even so, Mombasa, in quite new circumstances, in the 18th century reached the apogee of its power as an independent city-state. The architects of this achievement were the Mazrui, an Omani clan who had provided some of the imam's governors to Mombasa but who, because they were opposed to the Āl Bū Sa'īdis, did not long persist in their allegiance to Muscat. They owed their authority in Mombasa itself to an ability to hold the balance between the rival factions in the Swahili population and also to their ability peacefully to overcome all but one of their dynastic successions. In 1746 a Mazrui notable, 'Alī ibn Uthman al-Mazrui, overthrew an Omani force that had murdered his brother. Soon after he seized Pemba and, but for a family quarrel, might have won Zanzibar; his successor, Mas'ūd ibn Nāṣir, initiated a pattern of cooperation with Pate, maintained close links with inland Nyika peoples, and established Mazrui dominance from the Pangani River to Malindi.

Both Mombasa and Pate were disastrously defeated by Lamu in the battle of Shela, *c.* 1810. Pate's preeminence in the Lamu islands was destroyed, Mombasa's authority on the coast was diminished, and the way was open to Muscat's great intrusion into East African affairs. Lamu appealed to Oman for a garrison to assist it, to which Sayyid Sa'īd of Muscat very soon responded.

The rise of Zanzibar

The Āl Bū Sa'īdīs, who had captured Kilwa in 1785, maintained their principal footing upon the coast in Zanzibar, which had long held to its association with them. Thanks to the city's growing success, from the end of the 18th century onward, in turning itself into the main entrepôt for the trade in the area south of Mombasa, Zanzibar soon rivaled Mombasa as the focal point for the whole coastline. As such, it was both developed and used by Sayyid Sa'īd ibn Sulṭān of Oman as the base for his growing ambitions. Having won the succession to Muscat after an internecine struggle following his father's death in 1804, Sa'īd spent much of the next two decades establishing his authority there. (In this he was assisted by the British, who were much concerned to safeguard their route to India, which ran close to Muscat on its way past the Persian Gulf.) Then, in 1822, he wrested Pemba from Mazrui control and by 1824 had installed a Muscat garrison in Pate as well, thus bringing to an end the previous influence that the Mazrui had exercised.

Sensing the increasing threat from Muscat, the Mazrui appealed to the British for assistance. Though their application was formally denied, a British naval officer, Captain W.F. Owen, on his own initiative raised a British flag of protection over Mombasa in 1824. Since the British had no desire formally to extend their authority to East Africa at this time, let alone to break with their ally Sa'īd, it was hauled down in 1826. This gave Sa'īd his opportunity, and in 1828, 1829, and 1833 he mounted assaults upon Mombasa. But it was only when he successfully intervened in a dynastic dispute among the Mazrui, which followed on the death of a *liwali* in 1835, that he was able in 1837 to fasten his control over Mombasa and to topple the Mazrui from their position. His dominion along the whole coastline thus became assured, and after over a century's interval the East African littoral once more found itself dominated by a single outside power. Though this outcome owed much to the inability of the coastal towns to unite against an invader, it owed much as well to the striking personality of Sa'īd himself, to his investment in a navy, to his force of Baluchi soldiers (with which he supplemented his Omani levies), and to the support he received from the British.

It also stemmed from his intimate association with the major economic developments then taking place along the East African coast. These began with a marked growth in the previously marginal slave trade, particularly at first in the Kilwa region, more especially from 1780 to 1810 as a result of French demand for slaves in Mauritius and

Bourbon. This was succeeded by the discovery that cloves could be successfully grown on Zanzibar and by the development of flourishing plantations. British pressure on Sa'id to end the export of slaves to "Christian" markets came to fruition in 1822, when he reluctantly signed what became known as the Moresby Treaty. In the event, however, it made very little difference, either on the coast or in the interior, since slaves were being required in growing numbers for the plantations on both Zanzibar and Pemba and for export to the Persian Gulf and beyond.

Increasing commercial activity brought Sayyid Sa'id sufficient wealth to buy ships and pay troops. It also attracted to the East African coast migrant Indians, who became heavily involved in the country's economic expansion; and, together with the Arabs who were beginning to make profits from their clove plantations, Indians helped to finance the new upcountry trading caravans.

Trade with Europeans

The increased economic activity that centred upon the islands of Zanzibar and Pemba served to enhance the importance of the smaller towns that stood on the mainland opposite. It also attracted an influx of European traders, of which the most important were the Americans. They were the first Westerners to conclude a trade agreement with Sa'id (1833) and the first also to establish a consul at Zanzibar (1837). (Their prime achievement was to capture the cloth trade to East Africa—so that cheap cotton cloth thenceforth came to be known there as Americani.) The British followed with a trade agreement in 1839 and a consul in 1841. The French made similar provisions in 1844, and some Germans from the Hanseatic towns moved in at about the same time. British trade, however, never flourished and in fact died away; but by 1856 America and France were both making purchases in East Africa of more than $500,000 a year, while exports to India, particularly British India, were higher still. Some of the main items of trade, such as ivory, were traditional, but copal, sesame, cloves, cowries, hides, and coconut oil were also important. Because of this increased activity, Sa'id's economy in due course became less dependent upon the export of slaves, and he therefore showed himself more ready than he might otherwise have been to accept the so-called Hamerton Treaty of 1845, by which the export of slaves to his Arabian dominions was forbidden.

Since by this time the revenues from Sa'id's East African territories had overtaken those he received from Oman, it is understandable that in 1840 he should have transferred his own capital from Muscat to Zanzibar. At his death in 1856, Zanzibar was firmly established as the East African coast's main centre, from which major new incursions into the interior had begun to radiate extensively.

The interior before the colonial era. *The Stone Age.* The coast was never more than East Africa's fringe. Beyond the harsh *nyika,* or wilderness, which lay immediately inland and was nowhere pierced by a long, navigable river, thornbush country extended to the south, sometimes interspersed with pleasanter plains toward the centre, while to the north cooler forested highlands ran into harsher country. Westward lay the Great Rift Valley and, beyond, the regions of the great lakes whence the Nile ran northward through its usually impassable marshes.

Since there are no written records antedating the last century or so for this region, its history has to be deduced from often uncertain linguistic, cultural, and anthropological evidence; from oral traditions—where they are available, which at best is only for recent centuries; and from archaeological findings. Since investigations and analyses are still at a very early stage and since the first hypotheses have proved vulnerable to criticism, the statements that follow must be only tentative. Furthermore, all accounts of tribal migration must allow for innumerable short-run moves and may refer only to small—if important—minorities. They must also take account of probable interactions with other peoples en route and often, indeed, of extensive absorption. Above all, care must be exercised over anachronistic concepts of "tribe." Moreover, bolder categories such as Bantu are strictly only linguistic and must be treated with caution.

Two features of the pre-19th-century period may be stressed: first, although it seems to have been in this part of Africa that man first developed, in the three or four most recent millennia the key innovations in man's evolution seem to have occurred elsewhere; second, the extensive agricultural revolution in East Africa, which took place during this time, had the vital consequence that sizable populations grew up in areas of adequate rainfall, which could not be easily brushed aside by subsequent alien invaders.

Beginning of agriculture

During the earlier stages of the Stone Age down to about 50,000 BC, hand-ax industries were established in the Rift Valley areas of Kenya and of Tanzania (especially at Olduvai Gorge) and along the Kagera River in Uganda. During the Mesolithic period (thence to *c.* 10,000 BC), new stone-tool-making techniques evolved, and the use of fire was mastered. Spreading to other parts of East Africa, in the Neolithic period man clustered into specialized hunting and gathering communities from which may have developed some still-existing ways of life. The largest number of relevant sites is close to the homeland of the Hadzapi—the last contemporary hunters and gatherers—and to that of the Sandawe, who are physically and linguistically akin to the San (or Bushmen) of southern Africa. Remnants of other hunting and gathering communities—such as the Pygmies of western Uganda—or at least the memory of them, are found in many places. Latterly, they often lived in precisely those highland regions where agriculture and animal domestication in East Africa first occurred.

Food production and the keeping of cattle seem to have begun in the highland and Rift Valley regions of Kenya and of northern Tanzania in the 1st millennium BC and to have derived from Caucasoid peoples who were probably southern Cushites from Ethiopia. Some traces of these interlopers remain among, for example, the Iraqw of Tanzania, and it may be that the age-old systems of irrigation found throughout this region owe their origins to this period as well. Agriculture preceded the smelting of iron in these areas, and hunting and gathering continued to be important for the domestic economy. It looks as if in due course southern Cushites spread deep into what is now southern Tanzania, but, so far as has been ascertained, food production did not develop in the period BC elsewhere in Tanzania, nor in what is now Uganda.

The spread of ironworking and the Bantu migrations. It is still far from clear when and whence iron smelting spread to the East African interior. Certainly there was no swift or complete transfer from stone to iron. At Engaruka, for example, in that same region of the Rift Valley in northern Tanzania, a major Iron Age site, which was both an important and concentrated agricultural settlement using irrigation, seems to have been occupied for over a thousand years. Significantly, its styles of pottery do not seem to have been related to those that became widespread in the 1st millennium AD. It is a reasonable assumption that its inhabitants were Cushitic speakers, but it seems that its major period belongs to the middle of the 2nd millennium AD.

Spread of agriculture, ironworking, and Bantu speakers

The major occurrences of the 1st millennium AD involved the spread of agriculture—more particularly, the cultivation of the banana—to the remaining areas of East Africa. Simultaneously or perhaps previously went the spread of ironworking, and fairly certainly too the diffusion of Bantu languages—except in the core of the Cushitic wedge and to the north of an east–west line through Lake Kyoga. If, as seems probable, proto-Bantu languages had their origins in the eastern interior of West Africa, it does not seem inconceivable that over a lengthy period of time some of its speakers, probably carrying with them a knowledge of grain agriculture and conceivably a knowledge of ironworking, should have diffused along the tributaries of the Congo River to the savanna country south of the Congo forest into what is now the region of Shaba in Zaire. Nor does it seem inconceivable that the banana, originally an Indonesian plant particularly suitable in tropical conditions, should have spread to that same region up the Zambezi valley (certainly the Malayo-Polynesian influences in Madagascar in the 1st millennium AD are well attested in other respects). At all events, the linguistic and archaeological arguments for a fairly rapid eastward and northward expansion during the

1st millennium AD from the Shaba area now have wide acceptance. Bantu languages came to dominate most of this region (many Cushitic speakers in what is now Tanzania seem to have switched over to them or to have been eliminated). More varieties of banana developed in East Africa than anywhere else in the world. Ironworking was soon prevalent, and, where rainfall, soil nutrients, and the absence of the tsetse fly allowed, population growth increased decisively.

The early interlacustrine kingdoms. Sometime before the middle of the 2nd millennium AD, some of the most interesting developments were occurring in the interlacustrine area—*i.e.,* the region bounded by Lakes Victoria, Kyoga, Albert, Edward, and Tanganyika. Vague accounts of rulerships in various parts of this area date from the first half of the 2nd millennium AD, and it is at least possible that they existed—though they may well have been judicial arbitrators or ritual leaders rather than more strictly political figures. Whether they had their origins in roving Cushitic or Nilotic cattle keepers from the north or northeast—as has been variously suggested—is impossible to say, though some such explanation would not be difficult to believe. What seems certain is that around the middle of the present millennium a sudden cultural political climax was marked by a short-lived, though widely acknowledged, dynasty of Chwezi rulers.

Earthworks at Bigo

The Chwezi people are frequently associated with the great earthwork sites at Bigo, Mubende, Munsa, Kibengo, and Bugoma, in western Uganda. That at Mubende seems to have been a religious centre. The largest is at Bigo, where a ditch system, over six and a half miles long, some of it cut out of rock, encloses a large grazing area on a riverbank. It looks as if it comprised both a royal capital and a well-defended cattle enclosure. Its construction must certainly have required a considerable mobilization of labour—which, apart from indicating that it must have been the work of a substantial political power, would support the view that the distinction between cultivators and a pastoral aristocracy, which later became typical of this area, is of very long standing. Radioactive carbon dating suggests Bigo was occupied from the mid-14th to the early 16th century. This correlates with the evidence of oral tradition that around the turn of the 15th century the Chwezi were supplanted in the north by Luo rulers of the Bito clan (who provided the dynasties that ruled in Bunyoro, Koki, Buganda, and parts of Busoga) and that they were superseded to the south by various Hima rulers of the Hinda clan (in Ankole, Buhaya, Busubi, and around to the southeast of Lake Victoria). Under these, and the corresponding Nyiginya dynasty in Rwanda, powerful traditional rulerships among the interlacustrine Bantu persisted after the middle of the 20th century.

Their relatively common experience was reinforced in the aftermath of the Chwezi dynasty by the prevalence among them of a variety of (often commemorative) Chwezi religious movements. In some areas these took the form of spirit-possession cults; in others, pantheons of deities were developed. In various guises—sometimes in support of the existing political order, sometimes against it—they spread into Bunyoro, Buganda, Busoga, Ankole, Buha, Rwanda, Burundi, and even to Nyamwezi country, in what is now Tanzania. So extensive a diffusion of a basically common religious tradition in any other part of the East African interior before the much later arrival of Islām and Christianity was rare indeed.

The chieftainships of the southern savanna. In northwestern Tanzania, dynasties of a pre-Chwezi kind apparently spread from the interlacustrine area during the middle centuries of the present millennium. *Ntemi* (as the office was called) became prevalent among both the Sukuma and the Nyamwezi. They (the *ntemi*) were probably as much ritual leaders as political rulers; certainly they do not seem to have exercised before the 19th century a "state" authority that was characteristic of the later interlacustrine rulers. By about the 16th century there may have been an extension of this style of chieftainship southward into southwestern Tanzania. At all events, the chiefly groups among the Nyamwanga, the Nyika, the Safwa, the Ngonde, the Kinga, the Bena, the Pangwa, the Hehe, and the Sangu have common traditions of origin, and it seems clear that they are to be distinguished from their significantly different, matrilineal neighbours in southern Tanzania, Zambia, and Zaire. There also seem to have been secondary movements of *ntemi*-like institutions in the 18th century to Ugogo, Safwa, Kaguru, Kilimanjaro, and Usambara. At the same time, the development of chieftainships in these other areas of Tanzania may originally have occurred independently of influences from elsewhere.

Shungwaya

Northeastern Bantu. The spread of some Bantu to the northern coast of East Africa during the 1st millennium AD is supported by the memory of a settlement area named Shungwaya situated to the north of the Tana River. Shungwaya appears to have had its heyday as a Bantu settlement area between perhaps the 12th and the 15th centuries, after which it was subjected to a full-scale invasion of Cushitic-speaking Oromo peoples from the Horn of Africa. There is controversy as to whether the ancestors of the present Kamba and Kikuyu of Kenya were from Shungwaya, but it would seem that they probably broke away from there some time before the Oromo onslaught. It has been suggested, indeed, that the Kikuyu spread through their present territories from 1400 to 1800. The old Cushitic wedge checked them from spreading farther westward. This extended, as it would seem to have done for two or more millennia past, over both sides of the Kenyan and northern Tanzanian Rift Valley, but in the middle of the present millennium it was subjected to one of the multiple waves of invading Nilotic peoples—who were partly agriculturists and partly pastoralists—that moved into much of the northern and northwestern parts of East Africa.

The Nilotic migrations. The supersession of the Chwezi by Luo dynasties in the northern interlacustrine region at about the end of the 15th century resulted from the migrations of Nilotic peoples southward—in this instance, it has seemed, from a cradleland in what is now The Sudan.

For some 18 generations or so Bito rulers of Luo origin held sway over the kingdom of Bunyoro-Kitara, to the east of Lake Albert. Though at first their dominion seems to have been widely extended, they began to be rivaled in the 16th and 17th centuries by the rise of Buganda, under its ruler, or kabaka. Working on interior lines and based upon a particularly fertile region, Buganda developed a strength and cohesion that from the 18th century onward was to make it—with Rwanda—one of the two most formidable kingdoms of the region.

The Luo rulers and such followers as had accompanied them were soon fully absorbed into the Bantu population of these kingdoms. Immediately to the north (where the Bantu did not extend) there occurred the greatest independent expansion of the Luo peoples, who formed the Acholi; provided ruling groups for peoples to the west who came to be called Alur; and bred the Jopaluo and Jopadhola to the east and also the sizable Luo populations who, between the mid-16th and the mid-18th centuries, came to settle on the northern side of Winam Bay to the northeast of Lake Victoria and spread thereafter to its southern shore as well.

Over to the east, into the former Cushitic domain that centred upon the Rift Valley, there appears to have been, in about the middle of the present millennium, a great expansion of Kalenjin peoples. These Highland Nilotes (as distinguished from, among others, River–Lake Nilotes such as the Luo), seem to have absorbed most of the previous southern Cushites who remained there and also to have successfully held the core of this ancient wedge, if not its earlier dimensions, against further Bantu incursions. By 1700, however, a second expansion into this old protrusion was beginning. During the 18th century the Masai (Plains Nilotes, as they are now being called) spread over most of the area, until they came to be found as far south as Gogo country in central Tanzania. Already divided into pure-pastoralist and mixed-agriculturist subtribes, they were soon to be found to the east near Kilimanjaro. The earlier Kalenjin thus found themselves confined to the hillier country between the Rift Valley and Lake Victoria, where—constituting the Keyu (Elgeyo), the Suk

Expansion of the Masai

(Pokot), the Nandi, the Kipsikis, and the Tatoga of more recent times—they entered into a variety of interactions with their various Luo and Bantu neighbours. Farther to the north in the areas beyond Mount Elgon a shorter-run series of migrations by other Plains Nilotes was simultaneously taking place. First, in the 17th century, the Lango began moving southwestward (and became much affected by their River–Lake Nilotic neighbours, the Acholi). Then, in the 18th century, the Teso, Karamojong, and others began also to move in various southward directions.

It has been suggested that all these Highland and Plains Nilotic migrations were set off, both before and after the middle of the present millennium, by successive pressures from the Cushitic Oromo to the north. Like the Oromo, the Nilotic peoples lacked any firmly institutionalized political power, and their leaders were often less important than the elders of their clans. Indeed, Nilotes established "states" only where, as in the interlacustrine area over to the west, they came to rule other peoples who may very well have had traditions of rulership before they arrived. Such distinctions were to be of immense importance for the future.

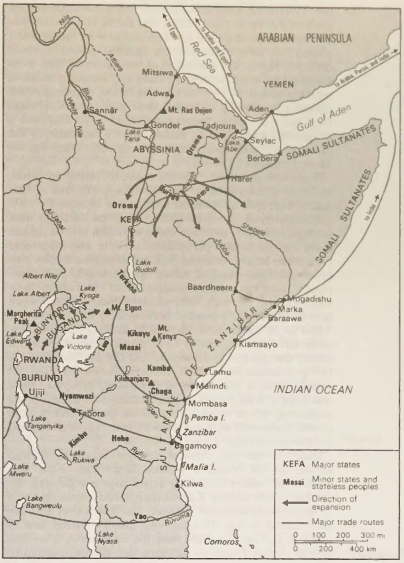
Pressure on the southern chieftainships. With the breakup of their main body, at the north end of Lake Nyasa, after 1845, some parties of Ngoni moved northward. Some Ngoni groups made their way to Songea; another struck north to Lake Victoria. They carried, to the area west of Lake Tanganyika, a new style of raiding and appear to have precipitated among such peoples as the Holoholo and the Ndendehule, the Sangu, and the Bena—in part, at least, as a safeguard against Ngoni raids—the creation of new political institutions, including more powerful rulerships. The most notable of such developments were, in the first place, among the Hehe, where, under Munyigumba and then on his death, in 1879, under his son Mkwawa, a powerful state was built up; then among the Nyamwezi, where between 1870 and 1884 the warrior chief Mirambo established a powerful personal rulership; and also among the Kimbu, where, between 1870 and 1884, Nyungu and his *ruga-rugas* (or bands of warriors) created a dominion that survived his death.

Nothing quite so striking occurred to the northeast. Conflict persisted between the smaller Chaga rulers on Kilimanjaro. In the mid-19th century, Kimweri enjoyed a considerable dominion in the region of the Usambara Mountains, but on his death in the 1860s his rulership disintegrated. Further such disintegrations occurred in the 19th century among the rulerships of Buha and Buzinza, at the south end of Lake Victoria, and among the Buhaya kingdoms on its eastern shore. The kingdom of Mpororo, to the west of Buhaya, had already broken up. By the 1890s its neighbour Nkore seems to have been in danger of disintegrating as well. Earlier in the century Toro, to its north, had broken away from Bunyoro, previously the most extensive of the kingdoms in this area, while fragmentation was almost endemic among the Busoga kingdoms to the east.

Decline of northeastern kingdoms

Rwanda and Buganda. But there were also growing points in the interlacustrine area, where one of the largest kingdoms, Rwanda, consolidated its rear by annexing Lake Kivu, then, in the aftermath of a succession war, swallowed the small kingdom of Gissaka, to its east. It failed to defeat Burundi, to the south, but under its mwami, or ruler, Kigeri IV (who reorganized its military forces) it extended its control by raiding to the north.

Its power was equaled in this region only by the kingdom of Buganda. Having annexed the large area of Buddu, to its southwest, in the late 18th century, Buganda thereafter generally refrained from any further territorial extensions. Its rulers steadily increased their authority at home by enhancing the power of appointed chiefs at the expense of the clan leaders, while abroad they preferred to make satellites rather than subjects of their neighbours. They had a good deal of success eastward in Busoga and southward along the western shore of Lake Victoria and around its southern rim. In the 1870s and '80s Buganda's protégés were on several occasions installed in petty rulerships in Busoga. In 1869 Bunyoro successfully survived Buganda interference in one of its succession conflicts (as Nkore did



Major states, peoples, and trade routes of eastern Africa, *c.* 1850.

in 1878) and indeed in the 1870s and '80s was renewing its strength. Bunyoro's improved position turned much on its new military formations, the *abarasura,* while Buganda's successful predation owed a good deal to its new military efforts under the *mujasi,* or military commander, as well as to the building of a formidable fleet of canoes.

The Luo and Masai. To the north and northeast the previous migrations of the Luo from west to east were followed in the 19th century by a new wave of migrations from east to west. The Lango, for example, further expanded in two southward and westward waves toward Lake Kyoga and toward the Victoria Nile, where they ran up against the Acholi. To their south the Teso and the Kumam were also moving west and south. A flourishing trading network developed around Lake Kyoga.

Activity was rife also among the pastoral peoples to the east. In about 1850 the Turkana began to migrate from a base west of Lake Rudolf. Southward stood the Masai, the warrior people of the plains and open plateaus north and south of the string of Rift Valley lakes west of Mount Kenya. From 1830 onward their various subtribes were engaged, under the auspices of their rival *laibons,* or ritual leaders—among whom Mbatian, who succeeded his father, Subet, in 1866, was the most famous—in a succession of internecine conflicts largely over cattle and grazing grounds. Their wars denuded the Laikipia and Uasin Gishu plateaus of their former Masai, the so-called Wakwavi, who, being deprived of their cattle, switched to agriculture. They also helped the Nandi, who, with the Uasin Gishu Masai now troubling them no more, took to raiding on their own account from a base between the Rift Valley and Lake Victoria. Under the leadership of their *laibon*-like *orkoiyots,* the Nandi and their kinfolk, the Kipsikis, were soon the new powers in the land. Some of their neighbours who lived in open country put up defense works against them—the Baluyia, to the west, for exam-

The Masai wars

ple, built mud walls around their villages—while others, such as the Teita, the Kamba, and the Kikuyu, who lived on higher ground and in forest country, were rather better placed and from their carefully guarded fastnesses could defy the Masai. On the edges of their country they even entered into some permanent trade and marriage relations with the Masai. Where the soil was fertile, moreover, such people considerably increased their populations. Though they had no chiefs, "prominent men" were accorded a recognized status among them, and by the close of the century some of these were fighting each other for local supremacy.

Trade with the coast. On the coast, following the death, in 1856, of Sayyid Sa'id, his erstwhile dominions in East Africa were split off from the imamate of Muscat. By 1873 the authority of the Āl Bū Sa'īdī sultans on Zanzibar itself became complete, although there were still revolts against them on the coast—particularly at Pate and Mombasa (where the Mazrui retained their preeminence despite successive defeats)—and at Kilwa, to the south. This arose chiefly from the sultan's acceptance of the further measures against the East African slave trade pressed upon him by the British consul at his court. By the 1860s some 7,000 or so slaves were being sold annually in the Zanzibar slave market, but in 1873 a treaty with the British closed the market at Zanzibar, and Sultan Barghash, by two proclamations in 1876, reduced the export from the mainland to a trickle. As it happened, however, there was then a final period of unprecedented slaving on the mainland, where the trade in slaves had generally been closely connected with the trade in ivory and the demand for porters was still considerable.

Trade in the East African interior began in African hands. In the southern regions Bisa, Yao, Fipa, and Nyamwezi traders were long active over a wide area. By the early 19th century Kamba traders had begun regularly to move northwestward between the Rift Valley and the sea. Indeed, it was Africans who usually arrived first to trade at the coast, rather than the Zanzibaris, who first moved inland. Zanzibari caravans had, however, begun to thrust inland before the end of the 18th century. Their main route thereafter struck immediately to the west and soon made Tabora their chief upcountry base. From there some traders went due west to Ujiji and across Lake Tanganyika to found, in the latter part of the 19th century, slave-based Arab states upon the Luapula and the upper reaches of the Congo. In these areas some of those who crossed the Nyasa–Tanganyika watershed (which was often approached from farther down the East African coast) were involved as well, while others went northwestward and captured the trade on the south and west sides of Lake Victoria. Here they were mostly kept out of Rwanda, but they were welcomed in both Buganda and Bunyoro and largely forestalled other traders who, after 1841, were thrusting up the Nile from Khartoum. They forestalled, too, the coastal traders moving inland from Mombasa, who seemed unable to establish themselves beyond Kilimanjaro on the south side of Lake Victoria. These Mombasa traders only captured the Kamba trade by first moving out beyond it to the west. By the 1880s, however, they were operating both in the Mount Kenya region and around Winam Bay and were even reaching north toward Lake Rudolf.

Zanzibari trade routes

The colonial era. Suggestions that he might at this time establish his dominion over the East African interior prompted Sultan Barghash to send a Baluchi force to Tabora, but the idea was never pursued. A comparable notion, however, led Khedive Ismā'īl Pasha of Egypt to appoint in 1869 the Englishman Samuel White Baker as governor of the Equatorial Province of the Sudan, so that Baker might carry the Egyptian flag to the East African lakes. Though Baker reached as far south as Bunyoro in 1872, he was soon obliged to leave. His successor, Charles George Gordon, proposed to circumvent both Bunyoro and Buganda by going straight up the Nile's banks. But Mutesa I, kabaka of Buganda, frustrated Gordon's efforts on the Nile, and by the early 1880s, with bankruptcy in Egypt and the Mahdist revolt in the Sudan, only remnants of the Egyptian enterprise remained.

The Egyptian incursion had been the climax to the search by many European explorers for the headwaters of the Nile—a quest that had obsessed the later years of the Scottish missionary David Livingstone and had prompted the discovery in 1858 of Lake Tanganyika by the English expedition of Richard Burton and John Hanning Speke. Speke returned first to discover Lake Victoria in 1858 and then with James Grant in 1862 became the first white man to set eyes on the source of the Nile, which Speke named Ripon Falls. By circumnavigating Lake Victoria 12 years later, Henry Morton Stanley stilled the controversy that had ensued in Europe over Speke's claim.

Missionary activity. The revelations of these explorers, the example of David Livingstone, concern in western Europe over the East African slave trade, and the Roman Catholic and evangelical fervour that existed there inspired the invasion of the East African interior by a motley collection of Christian missionary enterprises. Johann Ludwig Krapf and Johannes Rebmann of the Church Missionary Society,who had worked inland from Mombasa and had, in the 1840s and '50s, journeyed to the foothills of Mount Kenya and Kilimanjaro, were followed by a British Methodist mission. Roman Catholic missionaries reached Zanzibar in 1860 and settled at Bagamoyo in 1868. An Anglo-Catholic mission first tried to establish itself in the Shire highlands, then in 1864 transferred to Zanzibar. Anglican missionaries arriving in Buganda in the mid-1870s at the request of Kabaka Mutesa were soon followed by Catholic White Fathers—there and elsewhere on Zanzibar's Tabora route—while the London Missionary Society sent men both to Unyamwezi and to Lake Tanganyika.

The Church Missionary Society

There were, of course, a number of localized religious movements among the peoples of East Africa during the 19th century. These included the Mbari cult among the Nyakyusa, the Nyabingi in Rwanda, and the Yakany movement north of Mount Ruwenzori. None of them, however, spread in quite the way that the Chwezi movement had earlier. Islām, on the other hand—spread widely at the instance of the Zanzibari traders and long established on the coast—had secured a scattering of converts in the interior as in the key kingdom of Buganda.

This was the scene onto which Christian missionaries first entered. Although by 1885 there were nearly 300 of them in East Africa, they did not initially win many converts, and those they at first obtained came only from among freed slaves and refugees from local wars. After 1880, however, they made important conversions in Buganda, and by the end of the century Christianity was spreading in the Lake Victoria area over most of the region in which the Chwezi movement had previously percolated—and before very long over a much larger area as well.

Partition by Germany and Britain. Philanthropic, commercial, and eventually imperialist ventures followed these evangelical endeavours. Nothing, however, of great moment occurred until 1885, when a German, Carl Peters, riding a tide of diplomatic hostility between Germany and Britain in Europe, secured the grant of an imperial charter for his German East Africa Company. With this, the European scramble for Africa began. In east-central Africa the key occurrence was the Anglo-German Agreement of 1886, by which the two parties agreed that their spheres of influence in East Africa should be divided by a line running from south of Mombasa, then north of Kilimanjaro to a point on the eastern shore of Lake Victoria. This began the extraordinary process by which the territories and subsequently the nations of East Africa were blocked out first upon the maps far away in Europe and only later upon the ground in East Africa itself.

For histories of the East African states during the eras of colonial rule and independence, see below *The countries of East Africa: Kenya; Tanzania;* and *Uganda.*

For coverage of related topics in the *Macropædia* and *Micropædia,* see the *Propædia,* sections 943, 96/11, and 978, and the *Index.* (D.A.Lo./Ed.)

THE HORN OF AFRICA

The history of the Horn of Africa has largely been dominated by Ethiopia and has been characterized by strug-

Struggle between Muslim herdsmen and Christian farmers

gles between Muslim and other herdsmen and Christian farmers for resources and living space. The Christians mostly spoke Semitic languages and the Muslims Cushitic tongues. Although these languages were derived from the same Afro-Asiatic stock, the more apparent differences between the peoples often were excuses for war, which, by the end of the 20th century, was waged under the banner of nationalism and Marxism-Leninism.

Aksum. When the Ethiopian empire of Aksum emerged into the light of history at the end of the 1st century AD, it was as a trading state known throughout the Red Sea region. Its people spoke Ge'ez, a Semitic language, and they mostly worshiped Middle Eastern gods, although here and there a traditional African deity survived. Its port of Adulis received a continuous stream of merchants who offered textiles, glassware, tools, precious jewelry, copper, iron, and steel in return for ivory, tortoiseshell, rhinoceros horn, gold, silver, slaves, frankincense, and myrrh. Aksum, the capital, was five days' march from the coast onto the Tigray Plateau, from which position it dominated trade routes into the south and west, where the commodities originated.

By the 4th century Aksum had become a regional power and an ally of Constantinople, whose language and culture attracted the ruling elites. Sometime around 321 Emperor Ezana and the Aksumite court converted to the monophysitic Christianity—a belief that Christ had one nature that was both divine and human—of Alexandria's See of St. Mark. During the next 200 years Christianity penetrated the masses, as foreign and native-born monks proselytized the interior, building churches and establishing monasteries wherever they found pagan temples and shrines.

Through the first half of the 6th century Aksum was the most important state in the Red Sea–Indian Ocean region and even extended its power over the kingdom of Ḥimyarites on the Arabian Peninsula. In the Horn, Aksum dominated Welo, Tigray, Eritrea, and the important trade routes to and from the Sudan. The capital's stone buildings, monuments, churches, and 20,000 inhabitants were supported by tribute and taxes extracted from vassals and traders.

In 543 Abraha, the general in charge of Ḥimyar, rebelled and weakened Aksum's hold over South Arabia. This event marked the end of the empire's regional hegemony, allowed Persia to assume supremacy, and forced Constantinople into an overland trade route with India and Africa. Aksum's international trade diminished, a shift reflected in the debasement of the state's coins. The rise of Islām in Arabia a century later almost completely devastated Aksum, as Muslim sailors swept Ethiopian shipping from the sea-lanes.

Decline of Aksum

Aksum lost its economic vitality, and Adulis and other commercial centres withered. State revenues were greatly reduced, and the government could no longer maintain a standing army, a complex administration, and urban amenities. The culture associated with the outside world quickly became a memory, and Ethiopia learned to exist in local terms. The Christian state moved southward into the rich, grain-growing areas of the interior, where the rulers could sustain themselves. There, they and the local Cushitic-speaking population, the Agew, worked out a new political arrangement for Ethiopia.

The Somali. Meanwhile, another Cushitic people, the Somali, had separated themselves from the Oromo in what is now north-central Kenya. For their livelihood, they depended upon the one-humped Arabian camel, sheep, and goats. During the first centuries AD they migrated in a southeastern direction, finally following the Tana River to the Indian Ocean. They then turned north and peopled the entire Somali peninsula, coming into contact on the coast with Arab and Persian trading communities, from whom they took Islām and a mythological Arabian origin. By the 12th century the entire northern Somali coast was Islāmized, providing a basis for proselytism in the interior. But, as the Somali migration and Islām moved westward, they encountered a resurgent Christian Ethiopia.

The Solomonids. An amalgamated Christian state, led by Semitized Agew, had reappeared in the 12th century. This Zagwe dynasty gave way in the late 13th century to a dynasty that claimed descent from King Solomon and the Queen of Sheba, a genealogy providing the legitimacy and continuity so honoured in Ethiopia's subsequent national history. During the 14th and 15th centuries the Solomonid monarchs expanded their state southward and eastward. By then Muslims dominated Ethiopia's trade, which exited via Mitsiwa in Eritrea or through Seylac on the northern Somali coast.

The Solomonids permitted Muslim business activities in return for submission and taxes. In 1332 Ifat, a large Muslim polity with its port at Seylac, fed up with being a Christian vassal, declared a holy war against Ethiopia and invaded its territory, destroying churches and forcing conversions to Islām. The Ethiopian emperor, Amda Tseyon, fought back hard, routed the enemy, and carried the frontier of Christian power to the edge of the Shewan Plateau, overlooking the largely Muslim-inhabited Awash valley. One hundred years later, under Emperor Zara Yakob, the Solomonid empire extended its authority southward to modern Bale and Sidamo.

By then Ethiopia faced a challenge from Adal, Ifat's militant successor. Located in the semidesert Harer region, Adal employed highly mobile Somali and Afar cavalry, whose raiding Ethiopia could not control. Meanwhile, the Solomonid state had begun to decay, owing to succession problems and the sheer complexity of governing a large empire. The Muslims consequently stopped paying tribute and a percentage of their trading profits to the hated Christians. Thereafter, they grew stronger and more daring, responding in part to overpopulation among the Somali.

Aḥmad Grāñ

The distress was exploited by the charismatic Aḥmad ibn Ibrāhīm al-Ghāzī, known to the Ethiopians as Aḥmad Grāñ ("Aḥmad the Left-handed"). A pious, indeed rigorous, Muslim, Aḥmad railed against the secular nature of Adal, mobilized tribesmen to purify the state, and trained his enlistees to use the modern tactics and firearms recently introduced by the Ottomans into the Red Sea region. After he took over Adal about 1526, his refusal to pay tribute triggered a Solomonid invasion in 1527, which his army easily repulsed.

Aḥmad thereupon declared a holy war, led his men into Ethiopia, and won battle after battle, fragmenting the Solomonid state into its component parts. During 1531–32 the Muslims pushed northward, traversing the rich Amhara Plateau north of the Awash and destroying churches and monasteries. Aḥmad Grāñ built a civil administration composed of his own men, remnants of the pre-Solomonid ruling classes, and collaborators. By 1535 he headed a vast Islāmic empire stretching from Seylac to Mitsiwa on the coast and including much of the Ethiopian interior, but not the staunchly Christian mountain fastnesses.

There, in 1541, Emperor Galawdewos learned that 400 Portuguese musketeers had disembarked at Mitsiwa in response to pleas for assistance. Though they lost half their strength moving inland, their weapons and tactics inspired Galawdewos to exploit Ethiopia's difficult terrain by undertaking hit-and-run warfare. Aḥmad never knew where his adversaries would strike and therefore placed his forces in defensive positions, where they lost their mobility, while he and his personal guard acted as a rapidly deployed reserve. Encamped at Weyna Dega near Lake Tana, Aḥmad's unit was attacked on Feb. 21, 1543, by Galawdewos and a flying column. During the hard-fought battle Aḥmad was killed, and that single death ended the war.

Christian Ethiopia was reprieved, but at a great cost. The country had lost hundreds of thousands of lives, confidence in itself and its religion, and its store of capital. Unable to follow Europe into commercial and then industrial capitalism, Ethiopia rebuilt feudalism, because the state simply had to restore affordable administration. By the early 1550s Galawdewos had fashioned a reasonable facsimile of the high Solomonid empire. Muslims, especially in the border provinces of Ifat, Dawaro (in the modern Arsi region), and Bale, remained disaffected. Christian converts along the periphery of the heartland, south of the Blue Nile and the Awash River, chafed under

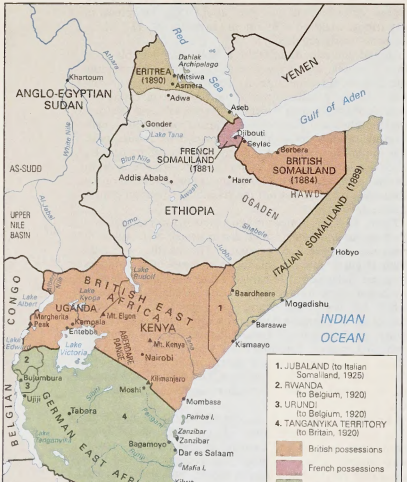
renewed exploitation, and the Judaized Falasha, to the north of Lake Tana, returned to their life of dispossession and economic marginalization. Finally, south of Lake Tana, in modern Gojam, Welega, Ilubabor, Kefa, Gamo Gofa, and Sidamo, a whole range of people remained tributary to the Christian kingdom. From among this last category emerged a new and more fundamental threat to old-fashioned Ethiopia.

Peoples tributary to the Ethiopians

Rise of the Oromo. The challenge came from the Oromo, a Cushitic-speaking pastoralist people whose original homeland was located on the Sidamo-Borena plain. From there, the related Afar and Somali peoples had hived off northeastward to the Red Sea coast, the Indian Ocean, and the Gulf of Aden, perhaps in some way causing the pressures that finally erupted in Aḥmad Grāñ's invasion of the Solomonid state. Some Oromo may have climbed onto the high Christian plateaus as early as the late 13th century, only to be repulsed. Garrisons established along the empire's periphery by Amda Tseyon and Zara Yakob were designed to keep the Oromo out, but when these defenses were destroyed during the war with Aḥmad Grāñ, the Oromo naturally resumed infiltrating.

The Oromo had an age-set form of government that changed every eight years, when a new warrior class sought its fortune by raiding and rustling in order to provide resources that the natural environment lacked. Every eight years, from the 1540s on, they advanced farther into the well-watered, fertile highlands—a sharp contrast with their arid bush country. Helped by their adversary's warweariness, demoralization, and depopulation, the Oromo invariably won territory after territory. By the beginning of the 17th century they had pushed northwestward into the modern regions of Arsi, Shewa, Welega, and Gojam and northeastward into Harerge and Welo, stopping only where they were blocked by forest, high population density, or effective mobilization of Christian forces.

Abyssinia. The Christians retreated into what may be called Abyssinia, an easily defensible, socially cohesive unit that included mostly Christian, Semitic-speaking peoples in a territory comprising most of Eritrea, Tigray, and Gonder and parts of Gojam, Shewa, and Welo. For the next two centuries Abyssinia defined the limits of Ethiopia's extent, but not its reach, for the Christian highlands received the hinterland's trade in transit to the Red Sea and the Nile valley. A complex caravan network linked Mitsiwa (now Massawa, Eritrea) on the Red Sea coast with the highlands of the interior. Gonder, the new capital, became a regional centre, doing business with the Sudanese cities of Sannār and Fazughli for slaves and gold, bought and paid for with coffee obtained from the Oromo-dominated lands. Demand for Ethiopian products increased considerably during the last quarter of the 17th century, as Yemen, a major trading partner on the Arabian Peninsula, sought increasing amounts of coffee for transshipment to Europe.



Eastern Africa as partitioned by the imperial powers, c. 1914.

Revival of the Ethiopian empire. By the late 19th century the northernmost Oromo had been assimilated into Christian culture, and Abyssinia's national unity had been restored after a century of feudal anarchy that ended with the accession of Yohannes IV in 1872. Yohannes forced the submission of Ethiopia's princes, repulsed Egyptian expansionism in 1875–76, pushed back Mahdist invasions in 1885–86, and limited the Italians to the Eritrean coast. Meanwhile, the ambitious King Menilek II of Shewa began a reconquest of Ethiopia's southern and eastern peripheries in order to acquire commodities to sell for the weapons and ammunition he would need in his fight for the Solomonid crown. Italian adventurers, scientists, and missionaries helped organize a route, outside imperial control, that took Shewan caravans to the coast, where Menilek's ivory, gold, hides, and furs could be sold for a sizable (and untaxed) profit.

The rise of Menilek

The economy of the Red Sea region had been stimulated by the opening of the Suez Canal, by the establishment of a British base in Aden, and by the opening of a French coaling station at Obock on the Afar coast. Britain sought to close off the Nile valley to the French by facilitating Rome's aspirations in the Horn. Thus, after 1885, Italy occupied coastal positions in Ethiopia and in southern Somalia. This limited the French to their mini-colony, leaving the British in control of ports in northern Somalia from which foodstuffs were exported to Aden. After Yohannes' death in March 1889, the Italians hoped to translate a cordial relationship with the new emperor, Menilek, into an Ethiopian empire.

On May 2, 1889, Menilek signed at Wichale (known as Ucciali to the Italians) a treaty of peace and amity with Italy. The Italians' famous mistranslation of Article XVII of the Treaty of Wichale provided them with an excuse to declare Ethiopia a protectorate. To Italy's dismay, the new emperor promptly wrote to the Great Powers, rejecting Rome's claim. Since neither France nor Russia accepted the new protectorate status, Ethiopia continued to acquire modern weapons from these countries through Obock. When, by 1894–95, Italy not only refused to rescind its declaration but also reinforced its army in Eritrea and invaded eastern Tigray, Menilek mobilized.

In late February 1896 an Ethiopian army of approximately 100,000 men was encamped at Adwa in Tigray, facing a much smaller enemy force some miles away. The Italians nevertheless attacked and were defeated on March 1, 1896, in what became known to Europeans as the Battle of Adowa (Adwa). Menilek immediately withdrew his hungry army southward with 1,800 prisoner-hostages, leaving Eritrea to Rome in the hope that peace with honour would be restored quickly. On Oct. 26, 1896, Italy signed the Treaty of Addis Ababa, conceding the unconditional abrogation of the Treaty of Wichale and recognizing Ethiopia's sovereign independence.

During the next decade, Menilek directed Ethiopia's return into the southern and western regions that had been abandoned in the 17th century. Most of the newly incorporated peoples there lived in segmented societies, practiced animal husbandry or cultivation with digging stick or hoe, followed traditional religions or Islām, and spoke non-Semitic languages. In practically every way but

Conquest of tributary peoples

skin colour, the northerners were aliens. Their superior weapons and more complex social organization gave them a material advantage, but they also were inspired by the idea that they were regaining lands that had once been part of the Christian state. Menilek and his soldiers believed that they were on a holy crusade to restore Ethiopia to its historic grandeur, but they did not realize that they were participating in Europe's "scramble for Africa" and that they were creating problems among nationalities that would afflict the Horn of Africa throughout the 20th century.

The birth of Somali nationalism. About 1900 the first of these problems erupted in Somali-inhabited regions, under the leadership of Sayyid Maxamed Cabdulle Xasan. The rebellion was directed at the British, Italians, and Ethiopians, whom Maxamed regarded equally as oppressors and infidels. Indeed, these powers admitted their collusion by collaborating militarily against the sayyid and his forces from 1901 to 1904, forcing him to sue for peace and to withdraw into a remote and unadministered area of Italian Somaliland. By 1908 he was again on the attack, this time causing a massive civil war, during which tens of thousands of Somali clansmen died. The Italians and the British chose not to intervene, preferring to let Somali kill Somali, and limited their activities to the coast. It was not until 1920 that British air power ran the sayyid to ground, forcing him to flee into the Ogaden, where he died on Dec. 21, 1920.

Maxamed pioneered the traditions of modern Somali nationalism, which combined Islām and anti-imperialism in a movement that sought to transcend clan divisions and make all Somali aware that they shared a common language, religion, way of life, and destiny. The Somali were further informed about their potential unity by, ironically, their Italian colonizers.

Italian rule. Despite the defeat at Adwa, Rome had not abandoned its dream of an Ethiopian empire. To this end, it worked hard at economic penetration but was invariably frustrated. More successful was its infiltration from Somalia into the adjacent Ogaden, where colonial troops seized strategic wells and posed as the protectors of Islām and the Somali people. By 1932 this advance alarmed Emperor Haile Selassie I, who was building a modern state in order to safeguard Ethiopia's independence.

Haile Selassie I

As regent to Empress Zauditu from 1916 to 1930, and afterward as monarch, Haile Selassie had worked to reform the economy, government, communications, and military. His success was recognized early, on Sept. 28, 1923, when Ethiopia entered the League of Nations. These achievements presaged a modern Ethiopian state that would block Rome's colonial plans and perhaps even undermine its position in the Horn of Africa. The potential threat of such a state, as well as considerations of European politics, led to the Italo-Ethiopian War, which began in the Ogaden, in December 1934, with a confrontation between Italian and Ethiopian soldiers at the water holes of Welwel.

Rome used Somalia and Eritrea as bases from which to launch its attack in October 1935. The issue was never in doubt; Haile Selassie had neither the armaments nor the disciplined troops necessary to fight the modern war that Italy mounted. In May 1936, after a terrible war that featured aerial bombardment and poison gas, he went into exile, and Italy proclaimed an East African empire consisting of Ethiopia, Eritrea, and its colony of Somalia.

The new regime, ignoring Ethiopia's traditional political organization, enlarged Eritrea to incorporate most of Tigray and placed the Ogaden in Somalia. In August 1940, after Rome had declared war against the Allies, the Italians marched north and occupied British Somaliland for seven months until dislodged by an Anglo-Ethiopian victory in the Horn of Africa. In 1942 and 1944 Anglo-Ethiopian treaties left the Ogaden under British rule for the duration of World War II, although Addis Ababa's sovereignty over the region was acknowledged. The British governed both Somalilands from a single city, Berbera, continuing the unification of the two territories.

Pan-Somalism. The Italians left Somaliland with an administrative infrastructure, communications, and towns, and the southern centres became incubators of pan-Somali ideas, which were quickly transmitted to their northern compatriots. The British allowed their subjects relative political freedom, and on May 13, 1943, the Somali Youth Club was formed in Mogadishu. Devoted to a concept of Somali unity that transcended ethnic considerations, the club quickly enrolled religious leaders, the gendarmerie, and the junior administration. By 1947, when it became the Somali Youth League, most of Somaliland's intelligentsia was devoted to pan-Somalism. This view was echoed in the British government's idea of Greater Somalia—a notion that was anathema to Ethiopia.

After his return to Addis Ababa in May 1941, Haile Selassie worked consistently to restore Ethiopia's sovereignty and to fend off British colonial encirclement and the isolation of his state. He regarded British activities in Somaliland as subversive and turned to the United States, which he concluded would be the dominant postwar power, to balance the geopolitical threat. American lend-lease and other assistance permitted Ethiopia to rebuff Britain and to secure the return of the Ogaden in 1948. The vision of Greater Somaliland, however, dominated Somali political programs in subsequent years.

The Ogaden returned to Ethiopia

Eritrean nationalism. Another fixed idea, that of Eritrean independence, also derived from the Italian years. Partisans here argued that Eritrea had evolved modern social and economic patterns and expectations from its colonial experience and, between 1941 and 1952, from the political freedoms allowed by the relatively liberal British military administration. While some Eritreans, especially Muslims and intellectuals, held these views in the 1940s, the idea of union with Ethiopia attracted the largely Christian population in the highlands—arguably the colony's majority. The Christians joined the Unionist Party, sponsored by the Ethiopian government, which simultaneously sought international support for regaining its coastal province. The Ethiopians were assisted by an international fact-finding commission that visited Eritrea in late 1948 and concluded that there was no national consciousness to nourish statehood and that its backward agriculture, crude industrial base, and poor natural resources would not sustain independence. The commission recommended some form of dependency—a decision ultimately referred to the United Nations, where the United States was the most influential power.

Washington was concerned about retaining control of a communications station near the Eritrean city of Asmera (now Asmara), which beamed intelligence information from the Middle East to the Pentagon, and it decided to support Ethiopia's claim to Eritrea in return for a formal base treaty. With U.S. leadership, the United Nations agreed to a federation of Ethiopia and Eritrea, which came into being in 1952. One year later, responding to growing Soviet influence in Egypt, Washington decided to provide Ethiopia military and economic aid. The United States subsequently became Ethiopia's main supplier of capital, expertise, and technology as well as military training, equipment, and munitions—a relationship that ultimately drove Somalia into an alliance with the Soviet Union.

Somalia irredenta. The Mogadishu government became independent on July 1, 1960. Its flag was dominated by a star, three points of which represented Djibouti, the Somali-inhabited northern region of Kenya, and the Ethiopian Ogaden. Together, these made up Somalia irredenta. In the Ogaden, young men organized themselves into clandestine fighting units, heeding Mogadishu's constant radio broadcasts to prepare for a war of liberation. In February 1963, the Ethiopian government sought to introduce a head tax to help sustain development efforts in the Ogaden. Somali nomads vigorously resisted the tax and rebelled, supported by the armed bands and then, in the fall of 1963, by Somalian troops. In November Mogadishu signed a military assistance pact with the Soviet Union, which undertook to equip a 20,000-man army. Shocked, the Ethiopians attacked Somalian border posts and adjacent towns in January 1964 and, after hard fighting, forced a cease-fire. Subsequent negotiations, however, were unable to resolve the differences between Somalia's goal of uniting all its compatriots and Ethiopia's need to retain its national integrity—as it was doing in Eritrea.

Efforts to absorb Eritrea

Cracks in the empire. From the Ethiopian federation's very inception, the imperial government had worked to transform Eritrea into an ordinary province. Courts, schools, and social services slowly became organs of the imperial regime; freedoms enunciated in the Eritrean constitution were suborned; political parties were suppressed; leading personalities were exiled; use of the Amharic language and other attributes of imperial culture were imposed on the population; the Eritrean flag was banned; and in 1960 the designation Eritrean government was changed to Eritrean administration. Furthermore, though Eritrea had received more development funds than any other region in Ethiopia, its towns and industries once sustained by the needs of the Italian colonial and British military regimes had difficulty competing in a national economy.

In July 1960 a group of mostly Muslim exiles in Cairo announced the establishment of the Eritrean Liberation Front (ELF). Its manifesto, which called for armed struggle to obtain Eritrea's rights, attracted the support of Syria, which eagerly offered military training for rebellion in a country tied to the United States and Israel. This largely Muslim movement received an infusion of young Christians after 1962, when, through Addis Ababa's manipulation, the Eritrean assembly voted to adopt the status of a governorate. The ELF now had sufficient strength to attack Eritrea's administrative and economic infrastructure. In December 1970 the imperial government declared a state of emergency in parts of Eritrea and stepped up counterinsurgency activities.

The government also used force in Bale and Sidamo between 1963 and 1970, putting down a rebellion among Oromo farmers and Somali herders against new land and animal taxes. This inevitably became involved with the politics of Somalia irredenta. By late 1966 rebels controlled both southern Bale and southeastern Sidamo and, at the same time, were attacking northern districts at will. It was not until the government sent in two army brigades and several squadrons of ground-attack jets that the rebellion was suppressed.

The Ethiopian government had proved unable to undertake the social and economic programs that could win the allegiance of the people. Force became the only tool of social control, partly because the emperor had grown reliant on the military but also because his government was inherently weak. For students in Addis Ababa, Haile Selassie was an agent of reaction, and his supporters were seen as exploiting Ethiopia for the benefit of the United States and its allies. They identified the monarchy and the ruling elites as the enemy of the people, pointing to the huge profits they made from sharecropping and other forms of capitalistic agriculture. Indeed, radiating from Addis Ababa was a zone of economic development that grew annually, dislocating traditional farmers. Peasant anxieties about land dispossession were loudly repeated by the students, who abhorred the realities of unequal economic growth and opted instead for the theoretical egalitarianism of unproved Marxist-Leninist models of development.

Problems in the countryside

Militarism in the Horn. Meanwhile, the Soviets were busily arming Somalia, which by 1970 had become the most militarized state per capita in the Horn of Africa, sustaining 20,000 troops. Ethiopia's armed forces remained between 45,000 and 50,000, with the portion of government expenditure devoted to the military actually declining from about 20 percent in 1970 to 14 percent by 1974. Addis Ababa thus managed to contain the Eritrean guerrillas and to keep the Somali in check with relatively modest outlays and was able to devote more of its resources to economic development programs. The imperial regime may have wanted to spend more on the military, but its chief arms supplier, the United States, had long before decided not to permit Ethiopia an offensive capability and therefore provided money and arms only for internal security and for frontier defense.

Rise of the Derg. By 1973 it was clear that the power behind the Ethiopian throne was the army. In early 1974 the government was unable or unwilling to respond to economic crises caused by the inflation of petroleum prices and by drought and famine in northern Ethiopia. When junior officers and other ranks went on strike over working conditions and inadequate supplies and equipment, the government resigned. Although a new cabinet was appointed, dissidents within the military organized into a central committee, called the Derg. This quickly became the real government as the emperor's men dissipated their energies in coping with a series of demonstrations.

Simultaneously, the army was being infused with fully developed Marxist-Leninist ideas by homegrown ideologues or by returnees from Europe and America. The Western dogma was swallowed whole by the more militant and socially conscious officers and men, whose agenda quickly became the abolition of the monarchy and the creation of a socialist state. On Sept. 12, 1974, Haile Selassie was deposed and, during the next year, most industry and all land were nationalized, new mass organizations were put in place, and programs were begun that could not be implemented effectively because the soldiers always sought a military solution to political problems. The situation in Eritrea therefore continued to deteriorate, and the west was abandoned to the insurgents, now dominated by the more secular Eritrean People's Liberation Front (EPLF). The EPLF took most of eastern Eritrea, leaving only the major centres in government hands.

Eritrean rebel groups

War in the Ogaden. In Addis Ababa, meanwhile, civilian opposition to the military government erupted in urban civil war. On Feb. 11, 1977, Mengistu Haile Mariam was named head of state and chairman of the ruling military council, and throughout 1977 anarchy reigned in the country as the military suppressed its civilian opponents. During this trauma the Somali chose to attack.

The Somalian president, Maxamed Siyaad Barre, was able to muster 35,000 regulars and 15,000 fighters of the Western Somali Liberation Front (WSLF). His forces began infiltrating into the Ogaden in May–June 1977, and overt warfare began in July. By September 1977 Mogadishu controlled 90 percent of the Ogaden and had followed retreating Ethiopian forces into non-Somali regions of Harerge, Bale, and Sidamo.

After watching Ethiopian events in 1975–76, the Soviet Union concluded that the revolution would lead to the establishment of an authentic Marxist-Leninist state and that, for geopolitical purposes, it was wise to transfer Soviet interests to Ethiopia. To this end, Moscow secretly promised the Derg military aid on condition that it renounce the alliance with the United States. Mengistu, believing that the Soviet Union's revolutionary history of national reconstruction was in keeping with Ethiopia's political goals, closed down the U.S. military mission and the communications centre in April 1977. In September, Moscow suspended all military aid to the aggressor, began openly to deliver weapons to Addis Ababa, and reassigned military advisers from Somalia to Ethiopia. This Soviet volte-face also gained Ethiopia important support from North Korea, which trained a People's Militia, and from Cuba and the People's Democratic Republic of Yemen, which provided infantry, pilots, and armoured units. By March 1978, Ethiopia and its allies regained control over the Ogaden.

Fall of military governments. Mengistu's government was unable to resolve the Eritrean problem, however, and expended large amounts of wealth and manpower on the conflict while rebellion spread to other parts of Ethiopia. Similarly, Siyaad proved unable to return the Ogaden to Somalian rule, and the people grew restive; in northern Somalia, rebels destroyed administrative centres and took over major towns. Both Ethiopia and Somalia had followed ruinous socialist policies of economic development, and they were unable to surmount droughts and famines that afflicted the Horn during the 1980s. In 1988 Siyaad and Mengistu agreed to withdraw their armies from possible confrontation in the Ogaden.

By 1989 Siyaad had refused serious political negotiations with his opponents, and fighting in Somalia spread southward and to Mogadishu. Amid increasing anarchy, the president fled in 1991, leaving Somalia to disintegrate into clan units.

Meanwhile, Mengistu refused to negotiate provincial autonomy, sparking the growth of ethnically based organizations. By 1987 the Tigre People's Liberation Front

(TPLF) controlled much of Tigray province. After a failed military coup in 1989, the TPLF advanced toward Shewa, attracting supporters from other areas. The TPLF joined with other forces to form the Ethiopian People's Revolutionary Democratic Front (EPRDF), which, with the EPLF, defeated Mengistu's forces throughout 1990 and 1991. Mengistu fled in May 1991, and the EPRDF began organizing an ethnically based government. The EPLF declared itself the de facto government of Eritrea and made moves toward independence. The intense upheaval, destitution, and fragmentation in the Horn of Africa put into question the future of political and territorial alignments.

(H.G.M.)

For coverage of related topics in the *Macropædia* and *Micropædia*, see the *Propædia*, sections 942, 96/11, and 978, and the *Index*.

THE COUNTRIES OF EAST AFRICA

Kenya

The Republic of Kenya (in Swahili, Jamhuri ya Kenya), a member of the Commonwealth of Nations, is one of the most strategically located countries of eastern Africa. Bisected horizontally by the equator and vertically by the 38th meridian (east), it is bordered on the north by Ethiopia and The Sudan, on the west by Uganda and Lake Victoria, on the south by Tanzania, and on the east by the Indian Ocean and Somalia. Kenya has an area of 224,961 square miles (582,646 square kilometres).

Diversity and unity

Kenya's present boundaries, the product of rivalries between colonial European powers, contain a number of ethnically diverse peoples who are independent and proud of their cultural heritage. Yet Kenyans are also acutely aware of the need to forge a strong national identity, of which cooperation is a basic ingredient. Since independence in 1963, the government has rallied the people under a national motto of "Harambee," or "Pulling together." Postcolonial development has been a mixture of pragmatic economic policy and of communal effort based on the principle of self-help. In this way the republic has tried to strengthen its traditional agricultural base as a foundation for industrialization. In addition, it continues to promote tourism, for the beauty and variety of Kenya's landscape, the pleasant and sunny climate, and the impressive dances and music of the Kenyan people are among the sure foundations that attract a growing stream of tourists to the country.

PHYSICAL AND HUMAN GEOGRAPHY

The land. *Relief.* The 38th meridian divides Kenya into two halves of striking diversity. While the eastern half slopes gently to the coral-backed seashore, the western portion rises more abruptly through a series of hills and plateaus to the Eastern Rift Valley, known in Kenya as the Central Rift. West of the Rift is a westward-sloping plateau, the lowest part of which is occupied by Lake Victoria. Within this basic division, Kenya is divided into the following geographic regions: the Lake Victoria basin, the Rift Valley and associated highlands, the eastern plateau forelands, the semiarid and arid areas of the north and south, and the coast.

The Lake Victoria basin is part of a plateau rising eastward from the lakeshore to the Central Rift highlands. The lower part, forming the lake basin proper, is itself a plateau area lying between 3,000 and 4,000 feet (900 and 1,200 metres) above sea level. The rolling grassland of this plateau is cut almost in half by the Kano Rift Valley, into which an arm of the lake known as Winam, or Kavirondo, Bay extends eastward for 50 miles (80 kilometres). The floor of the Kano Rift Valley, called the Kano Plain, merges north and south into highlands characterized by a number of extinct volcanoes. These include Mount Elgon, rising to 14,178 feet (4,321 metres) at the Ugandan border on the extreme north of the basin.

The Central Rift Valley splits the highlands region into two sections: the Mau Escarpment to the west and the Aberdare Range to the east. The valley itself is from 30 to 80 miles wide, and its floor rises from about 1,500 feet in the north around Lake Rudolf (in Kenya called Lake Turkana) to over 7,000 feet at Lake Naivasha but then drops to 2,000 feet at the Tanzanian border in the south. The floor of the rift is occupied by a chain of shallow lakes separated by extinct volcanoes. Lake Naivasha is the largest of these, the others including lakes Magadi, Nakuru, Bogoria, and Baringo. West of the valley, the diverse highland area runs from the thick lava block of the Mau Escarpment–Mount Tinderet complex northward to the Uasin Gishu Plateau. East of the Rift, the Aberdare Range rises to nearly 10,000 feet. The eastern highlands extend from the Ngong Hills and uplands bordering Tanzania northward to the Laikipia Plateau. Farther east they are linked by the Nyeri saddle to Mount Kenya, the country's highest peak, at 17,058 feet (5,199 metres). The relief of both highlands is complex and includes plains, deep valleys, and mountains. Important in the history and economic development of Kenya, the region was the focus of European settlement.

The Rift Valley highlands

The eastern plateau forelands, located just east of the Rift highlands, are a vast plateau of ancient rocks sloping gently to the coastal plain. They are a region of scattered hills and striking elevated formations, the most prominent being the Taita, Kasigau, Machakos, and Kitui hills. These hills form islands of more favourable climatic regimes and are surrounded by regions of greater aridity that make up the traditional famine areas.

The semiarid and arid areas in the north and northeast are part of a vast arid region extending from the Ugandan border to include Lake Rudolf and much of the plateau area between the Ethiopian and Kenyan highlands. (The area from Lake Magadi southward, though not as arid, shares the same characteristics.) There is scanty tree and grass cover here, but the areas of true desert are limited to the Chalbi Desert east of Lake Rudolf. The movement of people and livestock is strictly limited by the availability of water.

The coastal plain proper, which runs for about 250 miles along the Indian Ocean, is a narrow strip only about 10 miles wide in the south, but in the Tana River lowlands to the north it broadens to about 100 miles. Farther northeast, it merges into the lowlands of Somalia. The coast has excellent natural harbours, of which Mombasa is the best in East Africa.

Drainage. The major features of the drainage pattern of Kenya were created by the ancient crustal deformation of a great oval dome that arose in the west-central part of the country and created the Central Rift. This dome produced a primeval watershed from which rivers once drained eastward to the Indian Ocean and westward to the Congo system and the Atlantic Ocean. Still following this ancient pattern are the Tana and Galana rivers, which arise in the eastern highlands and flow roughly southeast to the Indian Ocean. West of the Central Rift, however, the major streams now drain into Lake Victoria. These include the Nzoia, Yala, Mara, and Nyando rivers. Between the eastern and western systems, the rifting of the dome's crust has created a complex pattern of internal streams that feed the major lakes.

Apart from the Tana River, most of the rivers of Kenya are short or ephemeral, disappearing in dry seasons. There are no major groundwater basins. Lake Victoria, extending over 26,828 square miles, is the largest lake in Africa and the second largest freshwater body in the world. Lake Rudolf, some 150 miles long and 20 miles wide, is the largest of the country's Rift Valley lakes. Other lakes are rather small and suffer large fluctuations in surface area.

Fertile lava deposits

Soils. In the Lake Victoria basin, lava deposits have produced fertile loam and sandy loam soils in the plateaus north and south of Winam Bay, while the volcanic pile of Mount Elgon is responsible for very fertile volcanic soils well-known for coffee and tea production. The

Central Rift Valley and associated highlands compose a distinct region of fertile, dark brown loams developed on younger volcanic deposits.

The most widespread soils in Kenya, however, are the sandy soils of the semiarid regions between the coast and the Rift highlands. To the north of the Rift are vast areas covered by red desert soils, mainly sandy loams. Because Kenya is very poorly forested, much of the soil is naturally exposed to widespread erosion. Erosion is stimulated as well by overgrazing and cultivation, especially in the arid and semiarid regions, and by encroachment of the limited forest areas.

Climate. Seasonal climatic changes are controlled by large-scale pressure systems of the western Indian Ocean and adjacent landmasses. From December to March, northeast winds predominate north of the equator and south to southeast winds south of the equator. These months are fairly dry, although rain may occur locally. The rainy season extends from late March to May, with air flowing from the east in both hemispheres. From June to August is a season of little rainfall in which southeast winds gradually prevail in the south and southwest winds blow north of the equator. October marks the beginning of a transition period ending with a return of the northeast and southeast winds.

In the Lake Victoria basin, mean annual rainfall varies from 40 inches (1,000 millimetres) around the lakeshore to well over 70 inches in the more elevated eastern areas. The lakeshore has excellent agricultural potential because it can expect 20 to 35 inches in 19 years out of 20. Daily maximum temperatures range from about 80° F (27° C) in July to 90° F (32° C) in October and February.

In the Central Rift Valley, mean annual temperatures decrease from about 84° F (29° C) in the north to just over 61° F (16° C) around Lakes Nakuru and Naivasha in the south. The bordering highlands are generally moderate, with mean annual temperatures ranging between 56° and 65° F (13° and 18° C). In general, the floor of the Rift Valley is dry, while the highland areas receive more than 30 inches of rain per year. In the Mau Escarpment, reliable rainfall and fertile soils form the basis of a thriving modern agriculture.

In most parts of the eastern plateau forelands, annual rainfall averages 20 to 30 inches, and agriculture is hampered by extremely uncertain rainfall. The semiarid and arid regions of northern, northeastern, and southern Kenya have very high temperatures and also suffer extremely erratic rainfall. Most places experience mean annual temperatures of 85° F (29° C) or more. Mean annual rainfall is only about 10 inches in the north and is less than 20 inches in the south.

The humid coast

Most parts of the coast experience mean annual temperatures in excess of 80° F (27° C) and high relative humidity year-round. From the humid coast, where mean annual rainfall is between 30 and 50 inches, precipitation decreases westward to about 20 inches per year. Only in the southern coast is rainfall reliable enough for prosperous agriculture.

Plant and animal life. In the highland area between 7,000 and 9,000 feet, the characteristic landscape consists of patches of evergreen forest separated by wide expanses of short grass. Where the forest has survived human encroachment, it includes economically valuable trees such as cedar (*Juniperus procera*) and podo wood (*Podocarpus milanjianus* and *P. gracilior*). Above the forest, a zone of bamboo extends to about 10,000 feet, beyond which there is mountain moorland bearing tree heaths, tree groundsel (a foundation timber of genus *Senecio*), and giant lobelia (a widely distributed herbaceous plant). East and west of the highlands, forests give way to low trees scattered through an even cover of short grass.

Semidesert regions below 3,000 feet give rise to baobab trees. In still drier areas of the north, desert scrub occurs, exposing the bare ground. The vegetation of the coastal region is basically savanna with patches of residual forests. While the northern coast still bears remnants of forests, years of human occupation in the south have virtually destroyed them.

Almost one-third of Kenya, particularly the western regions and the coastal belt, is infested with tsetse flies and mosquitoes. These insects are responsible for the prevalence of sleeping sickness and malaria in both the lake basin and the coastal region.

There is a close link between a region's vegetation and the differentiation and distribution of its wildlife. The highland rain forests support a variety of large wildlife dominated by elephant and rhinoceros, although these large animals have been much reduced by poaching and deforestation. Bushbuck, colobus monkeys, and occasional bushbabies are also found. In the bamboo zone are found varieties of the duiker. Predatory highland animals include lions, leopards, and wildcats. Birdlife is scanty.

The most prolific wildlife is found in the extensive grasslands between the forest zone and lower areas. The principal occupants here are varieties of ungulates, such as the hartebeest, wildebeest, zebra, and gazelle. Other members include waterbucks, impalas, elands, warthogs, and buffalo. Predators living on this grazing wildlife include the lion, spotted hyena, leopard, cheetah, and wild dog. Without the interference of the forest, birdlife is much richer here, and lakes and rivers are occasionally inhabited by hippopotamuses, crocodiles, and numerous fishes.

The coastal waters contain a wide variety of abundant fish life, including butterfly fish, angelfish, rock cods, barracuda, and spiny lobsters, while the coastal forest harbours suni antelopes, buffalo, and elephants.

In the thornbushes and thickets of the arid regions are elephants and rhinoceroses, lions and leopards, and giraffes, gerenuks, kudu, impalas, dik-diks, and lesser kudu. In the large rivers, hippopotamuses, crocodiles, and fishes may be found.

Traditional scattered settlements

Settlement patterns. Most of Kenya's population lives in scattered settlements, the location and concentration of which depend largely on climatic and soil conditions. Before colonization, virtually no villages or towns existed except along the coast, where pre-European urbanization was confined to fishing villages, Arab trading ports, and towns visited by dhows from the Arabian Peninsula and Asia. Mombasa, Lamu, and Malindi are among the few that expanded into modern towns. Other modern towns were established by Europeans at the coast and in the interior as administrative centres, mission stations, and markets.

Since independence, as urban areas have seen greater economic development, the migration from rural to urban areas has accelerated. In 1969 some 10 percent of the national population lived in urban areas of 1,000 or more people; by 1979 the urban population had grown to just under 15 percent. Along the coast, most of the urban population is found in the Mombasa complex, and in the interior the majority lives in the capital city of Nairobi.

The people. *Ethnicity, language, and religion.* The African peoples of Kenya, who account for about 98 percent of the total population, are divided into three main language groups. The largest of these is the Bantu group, which forms about two-thirds of the population and is largely concentrated in the southern third of the country. Bantu peoples occupying the fertile Central Rift highlands include the Kikuyu, Embu, Mbere, Kamba, and Tharaka. In the Lake Victoria basin they include the Luhya and Gusii.

The remainder of Kenyan Africans belong to the Nilotic and Cushitic language groups. The Nilotes, represented by the Luo, Kalenjin, Masai, and related peoples, make up about one-quarter of the total population. The rural Luo inhabit the lower parts of the western plateau draining into Lake Victoria, while the Kalenjin-speaking people occupy the higher parts of the plateau. The Masai are pastoral nomads in the southern region bordering Tanzania, and the related Turkana pursue the same occupation in the arid northwest.

The Cushitic-speaking peoples, who inhabit the arid and semiarid regions of the north and northeast, constitute only between 3 and 4 percent of Kenya's population. They are divided between the Somali, bordering Somalia, and the Oromo, bordering Ethiopia. The Cushitic peoples pursue a pastoral livelihood in areas that are subject to famine, drought, and desertification.

Apart from the African population, Kenya is home to various ethnic groups that immigrated during colonial rule from India and Pakistan. Referred to in Kenya as Asians, they are divided on the basis of religious affiliation into Hindus, Muslims, and Sikhs. Although many left after independence, a substantial number remain in urban areas such as Mombasa, Nairobi, and Kisumu.

European Kenyans, mostly British in origin, are the remnant of the farming and colonial population. At the time of independence, most Europeans emigrated to southern Africa, Europe, and elsewhere. Most of those remaining are to be found in the large urban centres of Nairobi and Mombasa.

Arabs (mostly the products of marriages between Arabs and Africans) live along the coast. Although all observe Islām, they are divided between the "old" Arabs, descended from Arabs who arrived before the 16th century, and the "true" Arabs, originating with the establishment of Arab hegemony in Zanzibar in the 19th century.

Swahili: the lingua franca

Although a wide variety of languages are spoken in Kenya, the lingua franca is Swahili. This multipurpose language, which evolved along the coast from elements of local Bantu languages, Arabic, Persian, Portuguese, Hindi, and English, is the language of local trade and is also used (along with English) as an official language in the National Assembly and the courts. In 1974 it became the official language, replacing English.

Kenya has no state religion. However, the majority of the Africans are members of the Roman Catholic, Anglican, and other Protestant churches. These religious affiliations are the outcome of early missionary activities, which assisted in the administrative pacification of the country during colonial times.

Population growth. In the late 20th century an accelerating population growth emerged as a serious constraint on social and economic development in Kenya. During the first quarter of the century, the total population was less than four million, and it was growing at a rate that would have doubled the population in 50 years. By the time of the 1948 census, the population had surpassed five million and was doubling every 30 years. By independence in 1963, the population had exceeded eight million and was doubling every 23 years, and by 1984 it had exceeded 19 million and the doubling time had been reduced to 18 years. The pressure of such a population explosion has been felt in limited employment opportunities, in the rising cost of education, health services, and food imports, and in the country's inability to generate the resources to build housing in both urban and rural areas.

The most important cause of this explosive growth has been a sharp fall in mortality rates—especially infant mortality—and a traditional preference for large families.

The economy. At the time of independence, Kenya's economy was characterized by a large traditional sector based on subsistence agriculture and the barter of goods, by a heavy dependence on foreign exchange for agricultural exports such as coffee and tea, and by a strong bond with the international economic system. Since 1963 the government has pursued a policy dedicated to a mixed economy of both privately owned and state-run enterprises. Most of Kenya's business is in private hands (with a great deal of investment by foreign firms), but the government also shapes the country's economic development through various regulatory powers and "parastatals," or enterprises that it partly or wholly owns.

The aim of this mixed economic policy has been to achieve economic growth and stability, to generate employment, and to maximize foreign earnings by maintaining the important agricultural exports while substituting domestically produced goods for goods that are traditionally imported. For a decade after independence, the policy showed great promise as rising wages, employment, and government revenue provided the means for expanding health services, education, and transportation and communication. But owing to setbacks, beginning with the rise of oil prices in 1973 and aggravated since then by periodic drought and accelerating population growth, Kenya's economy has proved unable to maintain a favourable balance of trade while addressing the problems of chronic poverty and growing unemployment.

Problems with economic growth

Agriculture. Although agriculture continues to dominate Kenya's economy, its share of the gross domestic product (GDP) declined from 42 percent in 1964 to 30 percent by 1987. Agriculture feeds the manufacturing sector with raw materials and supplies tax revenue and foreign exchange to support the rest of the economy.

Coffee and tea have continued to be the key foreign exchange earners. Other products that support external trade are pyrethrum, cotton, sisal, fruits, wattle bark, cashew nuts, and horticultural produce. For domestic consumption, the major farm products are corn (maize), potatoes, sugar, livestock, and poultry. Most of the country's livestock is raised in the arid regions by nomadic pastoral peoples who tend to keep them as a form of wealth rather than sell them for slaughter. Commercial stock raising is confined to small farms and large ranches, which raise animals for meat, hides, skins, and wool.

Despite the importance of agriculture to the economic well-being of the country, there are serious constraints on further expansion—in particular, the scarcity of water and the high cost of technological inputs.

Forestry. Although extremely important in the domestic economy, forestry is limited by the extent of forest land. Most of the total area of forest reserves is natural forest bush, bamboo, and grass; the remainder consists of planted softwoods, which now support a domestic paper industry. Forests are important in the conservation of Kenya's soil and water resources.

Fishing. Fish and other marine products represent a small but locally important portion of Kenya's natural resources. Freshwater fish from Lakes Victoria and Rudolf dominate fishery.

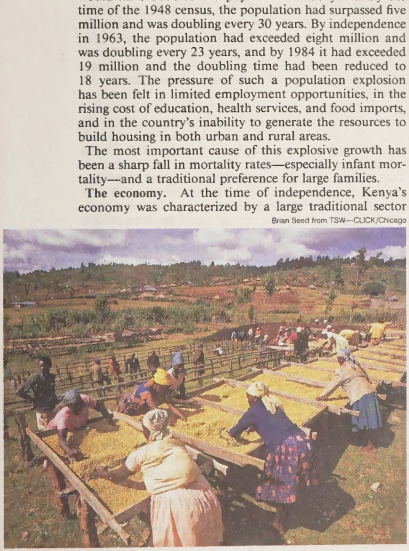
Mining. The mining industry is limited by the high cost of capital and by the limited market for Kenya's few mineral resources. Limestone deposits at the coast and in the interior are exploited for cement manufacture and agriculture. For the export trade, fluorite (used in metallurgy) is mined along the Kerio River in the north, and soda ash (used in glassmaking) is quarried at Lake Magadi.

Manufacturing. Manufacturing activity in Kenya is based on the processing of agricultural products, with agriculturally based industry accounting for about two-thirds of production. The country has a well-developed meat and meat-products industry and a thriving dairy industry. Production of sugar, textiles, paper, and leather has expanded greatly since independence. Exported agricultural products include processed fruits and extracts of pyrethrum (used in insecticides) and wattle bark (used in the tanning of leather).

Processing of agricultural products

Kenya's limited coal production supports several small steel-rolling mills. Petroleum products such as diesel and jet fuel, manufactured from imported crude at government-owned refineries near Mombasa, are important in

Brian Seed from TSW—CLICK/Chicago



Cooperative workers drying coffee on racks, Nyeri, Kenya.

both domestic and export markets. Also, engineering industries assemble motor vehicles and machinery from imported parts, and components are made from imported raw materials. (S.H.O.)

Tourism. Kenya is home to some of the rarest and most interesting species of wildlife in the world. Because of this, tourism is one of the country's major sources of foreign exchange, with visitors coming largely from countries of the European Union. Tourism revolves around a basic framework of some 20 national parks, game reserves, and game sanctuaries, where a wide variety of animals and cultural attractions can be enjoyed. The number of tourists began to vary annually, however, following a period of political unrest and attacks on tourists in the early 1990s.

Energy. Kenya's economic development depends largely on the development of energy. The emphasis since independence has been on the production of hydroelectricity, but this has limitations as an energy source for rural Kenyans, since the bulk of electricity is consumed by the two major urban centres of Nairobi and Mombasa. The largest hydroelectric plants are on the Tana River, though the considerable potential of the Turkwel River in the northwest is also under development. Geothermal resources in the Central Rift Valley have also been tapped since the early 1980s to generate electricity and now supply a significant amount of Nairobi's total needs.

Wood fuel, primarily used for domestic cooking, is a vital part of the rural economy, but deforestation threatens the supply. A tree-planting program has been initiated to establish in ecologically suitable areas quick-maturing indigenous and exotic species for the supply of farm households.

Finance. The Central Bank of Kenya, established by legislation in 1966, regulates the money supply, assists in the development of the monetary, credit, and banking system, acts as banker and financial adviser to the government, and grants short-term or seasonal loans. There are also more than 20 operating commercial banks, with hundreds of branches, sub-branches, agencies, and mobile units, which offer savings deposits and checking accounts. Financial mismanagement and corruption led a number of institutions to bankruptcy in the 1990s.

Trade. Coffee and tea comprise about half of Kenya's exports. Other agricultural products constitute another fourth, and petroleum products make up most of the remainder. The countries of the European Union are the primary export destinations, with the United Kingdom and Germany leading the list. Imports include machinery and transport equipment, crude petroleum, fertilizers, chemicals and pharmaceuticals, and manufactured goods. The main suppliers are the United Kingdom, Saudi Arabia, Japan, Germany, and the United States.

Transportation. The development, both before and after independence, of its excellent transportation infrastructure has largely made possible the emergence of Kenya as a modern and viable state.

The road system

Kenya's roads, the major form of transport linking the urban areas with their rural hinterlands, were developed in colonial times as a subsidiary of the railway line running from Mombasa to the western parts of the country. After independence the heavily utilized trunk and primary roads were upgraded from dirt to bitumen and gravel. With the expansion of this network came a rapid increase in freight traffic carrying goods within Kenya as well as to Tanzania, Uganda, The Sudan, and Ethiopia. This has severely damaged Kenyan roads, necessitating heavy expenditure in repairs.

Railways, the second most important mode of transport after roads, are operated by Kenya Railways. The main line runs northwest from Mombasa through Nairobi, Nakuru, and Eldoret to the Ugandan border. Major branch lines run from Nakuru to Kisumu on Lake Victoria and from Nairobi to Nanyuki near Mount Kenya. Passenger services constitute only a tiny share of railway business.

The strategic location of Kenya on the western shores of the Indian Ocean, with easy connections to different parts of Africa and the world, has greatly enhanced the role of Kenya's two international airports at Nairobi and Mombasa. A third international airport, at Eldoret, was completed in 1996. Kenya Airways, established as a state-owned company in 1977, was privatized in 1996. There are domestic airports at Kisumu and Malindi.

Mombasa, the principal port of Kenya, handles the bulk of the import and export traffic of Kenya, Uganda, and the northern mainland of Tanzania. The ports of Lamu and Malindi serve mainly the coastal trade and fisheries.

Administration and social conditions. *Government.* Kenya reached independence on Dec. 12, 1963, under a constitution that placed the prime minister at the head of a cabinet chosen by a bicameral National Assembly. A great deal of power was granted to assemblies elected in each of the country's regions, and multiparty contests were allowed. Since independence a series of amendments has abolished the regional assemblies in favour of provincial commissions appointed by the national government, made the National Assembly a unicameral body, proclaimed the Kenya African National Union (KANU) the only legal political party, and replaced the prime minister with an executive president who has the power to dismiss at will the attorney general and senior judges. The effect of these changes has been to establish the central government—in particular, the presidency—as the principal locus of political power in the country, although multiparty politics was once again allowed in 1991. The constitution guarantees a number of rights such as the freedoms of speech, assembly, and worship, but it also allows the president to detain without trial persons who have been deemed a threat to public security.

Membership in a political party is a requirement for anybody seeking election or appointment to public office except the presidency. Several candidates from the party are often allowed to contest a single office, but any candidate who receives more than 70 percent of the votes in a nominating election goes through the final election unopposed.

The executive branch

The executive branch consists of the president and the cabinet ministers, all of whom are selected by the president from the National Assembly. The president is the head of state and commander in chief of the armed forces. He is elected by direct popular vote to a term of five years and must also be a member of the National Assembly. Through the cabinet, the president controls the passage of legislation as well as the huge bureaucracies directing the economy and provincial affairs. Cabinet ministers frequently change portfolios, leaving the administration of the ministries to civil servants.

The National Assembly varies from 202 to 224 members. Most are elected for five-year terms by universal adult suffrage, with 12 members appointed by the president and two ex officio members—the speaker and the attorney general, of whom only the speaker has voting privileges.

The judiciary is headed by the chief justice and 11 puisne, or associate, justices of the High Court, which has full civil and criminal jurisdiction and rules on constitutional matters. The Court of Appeal, consisting of the chief justice and several associates, is the highest appeals court in the land. At lower levels are resident magistrates' and district magistrates' courts. Kenya's judicial system acknowledges the validity of Islāmic law and indigenous African customs in many personal areas such as marriage, divorce, and matters affecting dependents.

Local government consists of appointed provincial and district commissioners, elected county, municipal, and town councils, and elected township or municipal authorities. The provincial commissioners are responsible for education, transport, and health in their provinces, while the councils are concerned with services and public works funded by local taxes and grants from the central government.

Education. A single national educational system consists of three educational levels. At the first are eight years of primary education. These are followed by four years of secondary education, and the third level is represented by four years of university education. When primary schooling is completed, entrance to the secondary level is contingent upon passing an examination.

Primary and secondary enrollment have expanded markedly with the growing population, severely straining the government's ability to provide occupational training

for the slower-growing job market. As accommodations in government-built secondary schools are limited, community-built *harambee* secondary schools were developed, but promised government assistance for these schools has not always materialized. Public universities include Nairobi, Kenyatta, Moi (formerly Eldoret), and Egerton universities as well as Jomo Kenyatta College of Agriculture and Technology.

Health. With support from the World Health Organization and other foreign donors, vaccination programs have wiped out or reduced the incidence of such infectious diseases as smallpox and measles. Together with improved housing, education, sanitation, and nutrition, health-care programs drastically reduced death rates (especially the infant death rate) and raised life expectancy. However, AIDS has become a major disease in Kenya, and the death rate has again climbed. Life expectancy has also dropped, averaging about 49 years. In addition, the continued high incidence of malaria, gastroenteritis, dysentery, trachoma, and schistosomiasis illustrates the difficulty of eradicating mosquitoes and providing clean water, especially in the countryside.

Kenyatta National Hospital in Nairobi is the chief referral and teaching institution, and there are also provincial and district hospitals. In rural areas, health centres and dispensaries offer diagnostic services, obstetric care, and outpatient treatment. Rural health services suffer from lack of facilities, trained personnel, and medications.

Cultural life. There is a marked contrast between urban and rural culture. The cities are characterized by a more cosmopolitan population whose tastes reflect practices that combine the local with the global. Nairobi's night life, for instance, caters to youth interested in music that varies from American rhythm and blues, rap, and rock music to Congolese forms of rumba.

Rural life is oriented in two directions—toward the lifestyles of rural inhabitants, who still constitute the majority of Kenya's population, and toward foreign tourists who come to visit the many national parks and reserves, which are perhaps Kenya's greatest cultural legacy.

The Kenya National Theatre is incorporated in the Kenya Cultural Centre. The National Theatre School was founded in 1968 to provide professional training in theatrical techniques, which include the writing of plays by Kenyan authors and the performance of traditional music and dance. Music and dance play an integral role in social and religious life. Rhythm, all-important, is largely provided by the drum, supplemented by wind and stringed instruments. Swahili literature, both oral and written, is traditional in form and content. Contemporary novelists, including Ngugi wa Thiong'o and Mugo Gatheru, deal with the social frictions between traditional and modern society. Visual arts are largely confined to the mass production of wood sculpture for the tourist trade. Elimo Njau and Ronal Rankin are popular Kenyan painters.

(S.H.O./Ed.)

For statistical data on the land and people of Kenya, see the *Britannica World Data* section in the BRITANNICA BOOK OF THE YEAR.

HISTORY

Control of the interior. The earliest recorded history of the area now known as Kenya concerns the coastline, which for centuries has had trading relations with southern Arabia. There is little reliable information about the interior in the period up to the 19th century. During the 19th century Arab and Swahili caravans in search of ivory penetrated from Mombasa to Kilimanjaro and thence to Lake Victoria and beyond toward Mount Elgon, but this route was not so popular as were the caravan trails farther south, both because of the difficulty in crossing the desert country of the Taru Plain and because of the hostility of the warlike Masai tribe. The first Europeans to penetrate into the interior were the German agents of the Church Missionary Society, Johann Ludwig Krapf and Johannes Rebmann, who established a mission station at Rabai, a short distance inland from Mombasa in the territory of the Manyika tribe. In 1848 Rebmann became the first European to see Kilimanjaro, and in 1849 Krapf ventured still farther inland and saw Mount Kenya. These were isolated journeys, however, and more than 30 years elapsed before any other Europeans attempted to explore the country dominated by the Masai.

The first Europeans

The Masai and Kikuyu. The pastoral Masai probably moved into what is now central Kenya from the north in the mid-18th century. Their southward advance was checked about 1830 by the Hehe of what is now Tanzania, but their raiding parties continued to range far and wide and even reached the coast south of Mombasa in 1859. Under the spiritual direction of the *laibon* ("medicine man"), the Masai were organized for war. Great power was placed in the hands of the *moran* ("warrior"), so that, although the Masai were not particularly numerous, they were able to dominate a considerable region and to terrorize their neighbours. Their chief victims were Bantu agriculturalists, who could offer little effective resistance to the unexpected raids of these nomadic people. The Nandi, who inhabited the escarpment to the west of the Masai, were equally warlike and remained relatively undisturbed by their predatory neighbours. On the eastern slopes of Kilimanjaro, the Taveta were able to take refuge in the forest, and, still farther east, the Teita used the natural strongholds provided by their mountainous homeland to resist the Masai raiders.

The Kikuyu, a much larger tribe and near neighbours of the Masai, also looked to the mountains and forests for protection against Masai war parties. The Kikuyu had expanded northward, westward, and southward from their territory in the Fort Hall area of present-day Central province, cutting down the forests to provide themselves with agricultural land. Toward the end of the 19th century, however, their advance had reached the limits imposed by the presence of the Masai to the north and south and by the upper slopes of the Aberdare Range to the west. To the south, in the region of Ngong, they were protected by a narrow belt of forest through which the Masai could penetrate only at risk of damage from traps and ambushes. The Kikuyu were organized in clans and by districts, and, although war leaders arose to help in the defense of the tribe against the Masai, the lack of a central form of government made them vulnerable to the raids of their neighbours.

Decline of the Masai

In the 1890s famine and smallpox compelled the Kikuyu to withdraw northward, vacating much of the land in what is now Kiambu district. The Masai, too, were passing through a difficult period. An outbreak of disease, probably pleuropneumonia, had attacked their cattle in 1883, and in 1889–90 epidemics of rinderpest and smallpox further undermined their wealth and power. Simultaneously, the tribe was split on the death of its great *laibon,* Mbatian, and it was some time before his younger son, Lenana, was able to restore order. The legend of the Masai menace lingered on, but their power never revived, for their decline coincided with the arrival of European traders and administrators who took control of the country that the Masai had formerly ravaged but never ruled.

The British East Africa Company. In 1883 Joseph Thomson, an explorer, became the first British traveler to pass through the Masai country, but it was the British East Africa Company that undertook the opening up of the land to the west of Mombasa. German interest in East Africa had resulted in November 1886 in the delimitation of the grandiose but inadequately authenticated territorial claims of the sultan of Zanzibar. After recognizing the sultan's authority over a 10-mile-wide coastal strip between the Ruvuma and Tana rivers, Germany, Britain, and France had agreed to divide the hinterland into British and German spheres of influence. The British took the area north of a line running from the mouth of the Umba River, opposite Pemba Island, and skirting north of Kilimanjaro to a point where latitude 1° S cuts the eastern shore of Lake Victoria. The German sphere was to the south of that line. In 1887 Sir William Mackinnon and the British East Africa Association (later Company) accepted a concession of the sultan's territory on the mainland for a 50-year period, subsequently amended to a grant in perpetuity. The British government, reluctant to become involved in the

administration of East Africa, agreed in 1888 to a petition from the association asking for a charter of incorporation authorizing the acceptance of existing and future grants and concessions for the administration and development of the British sphere.

The financial resources of the new Imperial British East Africa Company were scarcely adequate for any large-scale development of East Africa. Becoming involved in civil war in Uganda, the company soon found itself unable to maintain its high level of expenditure and was forced to limit its activities to the region nearer the coast. In 1894 the British government declared a protectorate over Buganda and, in the following year, in order that it might itself guard the line of communication between the coast and Buganda, called upon the company to surrender its charter and concession in return for £250,000 compensation. The East Africa Protectorate was then proclaimed, and Sir Arthur Hardinge became the first commissioner. The lack of importance attached to the new protectorate by the British government was indicated by the fact that Hardinge continued to maintain his headquarters in Zanzibar, where he already fulfilled the duties of consul general.

Opposition to colonial rule

The East Africa Protectorate. The early years of the new administration were largely taken up in asserting its authority over tribes that did not understand the change that had taken place. This could scarcely be said of the chiefs of the Mazrui family along the coast, who actively resisted the usurpation of their authority by the British administrators. The Kikuyu and the Kamba, however, simply opposed what appeared to them to be an unwarranted invasion of their territory, and in 1896 and 1897 small military expeditions had to be sent against them by the new administration. Farther west the Nandi did not accept their new overlords until 1905, after a series of military columns had ranged through their territory with varying success. The Masai alone among the more important tribes offered no resistance to British authority.

The lack of communication and the limited financial resources available locally or granted by an apathetic British government meant that orderly administration was only slowly extended into the more remote areas of the protectorate. The absence of any strong indigenous forms of political organization that might have been employed as agents of the new protectorate also delayed progress. The introduction of direct rule through the agency of European administrative officers was not, therefore, so much a question of policy as a matter of necessity.

The Uganda railway and European settlement. At the beginning of the 20th century, the importance of the East Africa Protectorate was thought to lie in its value as a corridor to the fertile country around Lake Victoria. This view was further strengthened by the fact that the Kedong River marked the western boundary of the protectorate, so that the fine highland pastures to the west of the Rift Valley still formed part of the Uganda protectorate. Consequently, although regulations were issued in 1897 authorizing the lease to Europeans of land that was not cultivated or regularly occupied by Africans, there were few requests for concessions.

Two factors changed this negative attitude toward the protectorate. The first was the construction of a railway from the coast to Lake Victoria, and the second was the transfer of the western highlands from Uganda to the East Africa Protectorate in 1902. The need for a railway to promote the development of East Africa's economic resources had been emphasized by the Imperial British East Africa Company, and the British prime minister, Lord Salisbury, had favoured the idea, mainly on strategic grounds. There had, however, been strong opposition from anti-imperialists in Britain, and in 1891–92 a preliminary survey only had been carried out by Captain J.R.L. Macdonald. But the declaration of the Uganda protectorate strengthened Lord Salisbury's case, and work began on the construction of the railway at Mombasa in December 1895. The first locomotive reached Kisumu on Lake Victoria in December 1901, and the line was completed in 1903. By this time it had cost the British treasury £5,317,000, and, because of the alteration of the boundary between the two protectorates, it no longer extended to Uganda. To the burden of financing the administration of the East Africa Protectorate was now added the responsibility of making the railway pay. The proceeds of traffic to and from Uganda would still be available to the protectorate administration, but this did not remove the need to develop the resources of the territory to the utmost and as speedily as possibly. In these circumstances the decision to encourage extensive European settlement was not surprising. The African peoples were not organized in such a way as to suggest that African agriculture, even under European guidance, would produce the swift returns that the economy of the protectorate demanded. Moreover, the transfer of the highlands from Uganda had placed at the disposal of the administration large tracts of land apparently suited to large-scale farming, virtually uninhabited and situated in a region climatically attractive to Europeans. Even around Nairobi itself, which because of its central position was promoted early in the century from its former status as an important railway stores depot to be the administrative capital of the protectorate, there was apparently a considerable area of fertile, uninhabited land. For many reasons, therefore, the decision of Sir Charles Eliot, who had become commissioner for the protectorate in 1901, to invite European settlers to East Africa appeared sound. Eliot was both a scholar and a humanitarian, so that his policy was not adopted without a full consideration of its likely effects upon the indigenous African population. To make up for the lack of interest of Englishmen in East Africa, Eliot invited South Africans to settle in the protectorate. Since the response was both quicker and greater than had been expected, a hurried survey, based upon inadequate information, was carried out. As a result, small areas of land in Kiambu district, which had formerly been occupied by Africans and which the Kikuyu regarded as part of their legitimate area of expansion, were allocated to white settlers. The area involved was not large, but it served as a justification for all future African criticisms of European settlement.

Encouraging European settlement

As the settled area was extended during the first decade of the 20th century into the Rift Valley and on to the highlands beyond, the problem of labour supplies also arose. Much of this land was either uninhabited or supported only a very lightly scattered population. Having been invited to East Africa by the government, the settlers expected that an adequate labour supply would be made available. Because few Africans appeared to be completely occupied all the year round in work on their own farms, to the white settlers there seemed no reason why they should not be invited and, if necessary, compelled to offer their services to European farmers. The Africans, however, accustomed to subsistence farming and requiring only a low standard of material comfort, could see no reason why they should work away from home. Successive commissioners and governors responded in varying degrees to the settlers' demand for the introduction of some form of compulsion to ensure an adequate labour supply, but it was not until immediately after World War I, and largely as a result of a public outcry in Britain, that compulsory labour on either public or private projects was strictly forbidden.

Tribal reserves

Another cause of African discontent was the decision taken soon after Eliot's departure in 1904 to confine the native population to reserves. Although this plan was intended to protect the Africans from excessive competition as well as to leave as much land as possible available for scientific development by European farmers, the fact that the reserves were not gazetted for many years aroused a profound sense of unease among the Africans, for whom land was the only form of social security.

Owing to the Africans' reluctance to work on the construction of the railway, thousands of Indian coolies were brought into the protectorate; but most of these returned to India after their contracts were completed. The opening of the railway had, however, encouraged Indian traders who had formerly lived nearer the coast to penetrate far into the interior, even ahead of the administration, and these men made a valuable contribution to the country's economic development. Other Indians hoped to obtain land, but European settlers strongly resisted any idea of

their encroachment into the highland areas. Conscious of the great efforts they had put into the development of a virgin land and anxious to establish a truly British colony in East Africa, the Europeans consistently opposed the Indians' claim to equality in both political and economic fields.

Before the outbreak of war in 1914, unofficial participation in political affairs was mainly limited to the creation of pressure groups, the most prominent being the Convention of Associations, which had developed in 1911 from earlier European settler organizations. In 1905 an Executive Council had been appointed, and in 1907 the first Legislative Council was convened. Although the latter body consisted at first of only two nominated unofficial members in addition to the six official members, the unofficial members were not slow to uphold the white settlers' cause. The transfer of the control of the protectorate from the Foreign Office to the Colonial Office on April 1, 1905, did not give the settlers the increased responsibility for which they had hoped, and in 1913 they launched a campaign to elect their own representatives to the Legislative Council. The outbreak of World War I put a check upon their hopes, but the prominent part played by elected representatives in the activities of the War Council appointed in September 1915 offered a fair prospect of the early satisfaction of the settlers' claims to elected representation in the country's legislature.

World War I and its aftermath. The threat to the protectorate, and in particular to the vital Uganda Railway, from German East Africa never materialized, and early in 1916 the British advance into German territory finally removed the possibility of further military action on British soil in East Africa. Nevertheless, the war had serious effects upon the protectorate's economy. Most of the white settlers hastened to join the armed forces, leaving their farms to be looked after by their wives or else to revert to wilderness. Thus, at a crucial stage in the agricultural development of the territory, momentum was lost. Immediately after the war an attempt was made to stage a quick revival by the introduction of a "soldier settler" scheme, but the hopes of prosperity encouraged by the postwar demand for primary produce received a severe setback in the early 1920s, when the world depression brought bankruptcy to many of those who had started out with inadequate capital or who had relied upon credit from the banks. The achievement of stability was further delayed by the replacement of the rupee currency, first by a florin currency and shortly afterward, before the florin had been fully accepted, by East African shillings.

Economic depression

The courage, if not the confidence, of the settlers remained unshaken, and by the mid-1920s the country's economy had wholly revived. Although the depression of the early 1930s brought further difficulties to East Africa, the foundations of prosperity had been still further strengthened by a great improvement in rail communications. Already in 1913 railway extensions had been opened to Thika and to the soda deposits at Lake Magadi, and during the war a link was made between the main line and the German railway system to the south.

The most important postwar project was the building of a new extension of the main line across the Uasin Gishu Plateau to tap the agricultural wealth of the highlands, and thence to Uganda in order to provide a ready means of egress for the cotton crops of that protectorate. The line was eventually completed from Nakuru to Jinja in January 1928 and was carried on to Kampala, which it reached in January 1931.

Kenya Colony. *Political movements.* Political considerations then began to usurp the place of prominence formerly occupied by economic problems. In 1920 the status of the East Africa Protectorate was changed to that of Kenya Colony, named after the highest mountain in the country, and the coastal strip leased from the sultan of Zanzibar became the Kenya Protectorate. In the previous year the white settlers had been permitted for the first time to elect members to the Legislative Council, and the Indians began to call for equal treatment. In 1920 they rejected an offer of two elected members on the ground that this would mean that they were not so well represented as the Europeans, and a fierce political struggle broke out between the white settlers and their Indian rivals. In 1923, however, the colonial secretary, the Duke of Devonshire, after hearing representations from both Europeans and Asians, issued a White Paper in which he declared that African interests must be paramount, although minorities would not lose their rights. This pronouncement marked the tentative beginning of an era of "trusteeship" for the African peoples of Kenya. In 1927 the dispute between Europeans and Indians virtually ceased when the Indians accepted 5 seats in the Legislative Council. The Europeans were to have 11 elected representatives.

Trusteeship did not immediately result in any great improvement in the social conditions of the African population. Little provision for education was made by the government, and nearly all the schools for Africans were supplied by missionaries. Africans, too, found no place in the country's legislature, their interests being represented by the official members of the council and by a European unofficial member, usually a missionary. With the example of the Convention of Associations before them, however, the Africans developed their own pressure groups. Among the Kikuyu, who supplied a considerable proportion of the labour force on the European farms and whose proximity to Nairobi brought many of them into regular contact with Europeans, organizations to represent the tribe's grievances began to spring up quickly in the 1920s and 1930s. The most prominent was the Young Kikuyu Association, founded in 1921, which in 1925 was renamed the Kikuyu Central Association. In demanding African representation in the legislature, the association was in advance not only of the government but also of most of the members of the tribe. It won support among the Kikuyu, however, when it complained about low wages, the prohibition of coffee growing by Africans, and the condemnation by Christian missionaries of such tribal practices as clitoridectomy. The association never represented the tribe as a whole, though, because its members were mainly young men whom the chiefs did not trust. For this reason, too, the European administration tended to look with disfavour upon its activities. Attempts to win the support of other tribes failed owing to their unwillingness to accept Kikuyu leadership.

African political groups

Another cause for political concern was the proposal, first made toward the end of World War I, to introduce some form of closer union between Uganda, Kenya, and the former German possessions in East Africa. A number of commissions visited East Africa between 1924 and 1929 to investigate means of implementing the plan, but in the face of local opposition the idea was dropped early in the 1930s. At the outset the white settlers of Kenya opposed closer union with the other territories through fear of African domination. In view of the apparent determination of the British government, however, they agreed in the later 1920s to a compromise that would protect their political status in Kenya. In the 1930s they went further and actively supported union with Tanganyika as a protection against Germany's claims to its former overseas dependencies.

World War II to independence. The outbreak of World War II again checked development, and the entry of Italy into the war for a time threatened the security of Kenya's northern border with Ethiopia and Somaliland. The Italians were soon defeated, however, and troops from Kenya then took part in campaigns in many parts of the world. In Kenya itself the main objective was to achieve the highest possible degree of economic self-sufficiency. Political progress was not neglected, and in 1944 Kenya became the first East African territory to include an African in its Legislative Council. The number was increased to two in 1946, four in 1948, and eight in 1951. All were appointed by the governor from a list of names submitted to him by local governments. In 1948 the East Africa High Commission was set up to administer certain services of common benefit to Kenya, Uganda, and Tanganyika. With headquarters in Nairobi, it consisted of the governors of the three territories, with the governor of Kenya serving as chairman.

Both political progress and the large-scale economic de-

The Mau Mau Rebellion

velopment program initiated soon after the war received a severe setback in the 1950s with the outbreak of the Mau Mau Rebellion, which necessitated the proclamation of a state of emergency in October 1952. The rising was limited to the Kikuyu tribe and was directed against the presence of Europeans in Kenya and their ownership of land. Nonetheless, during the years of violence that followed, it was those Kikuyu who failed to support the Mau Mau secret society who suffered most heavily. Jomo Kenyatta, charged with directing the Mau Mau movement and sentenced with five associates to seven years' imprisonment in 1953, was released in 1959 and confined to restrictive residence. A considerable military campaign was needed to destroy the rebellion, and years of rehabilitation were required before order was restored. The emergency was brought to an end early in 1960.

Numerous economic and social changes resulted either directly or indirectly from the Mau Mau rising. One of the most important was the centralization of the Kikuyu in large villages, which was accompanied by a land consolidation program. This plan was also extended to Nyanza region near Lake Victoria. Considerable sums of money were set aside to encourage higher standards of African cultivation, and coffee became one of the principal crops. Throughout the years of uncertainty, foreign investment in Kenya continued. Limited industrial development took place with agricultural expansion.

In spite of the upheaval, successive colonial secretaries persisted in their attempt to promote political progress among the African population. In March 1957 elections were held on a qualitative franchise for the eight African seats in the legislature. In 1958 there were further proposals for an increase in African membership of the council, and a council of state was set up to scrutinize legislation that might involve discriminatory measures against certain sections of the people. The former rivalry between Europeans and Indians was now giving way to a rivalry between Europeans and Africans, and further constitutional advancement pointed the way to majority rule by Africans in Kenya. As a result of a conference held in London in 1960, Africans became the majority in the Legislative Council and obtained 4 of the 10 unofficial portfolios in the Council of Ministers.

To meet this challenge, two African political parties were formed: the Kenya African National Union (KANU), led by Tom Mboya and later by Jomo Kenyatta; and the Kenya African Democratic Union (KADU), led by Ronald Ngala. The former drew most of its support from the Kikuyu and Luo tribes and favoured a strong, centralized government; the latter originated among the smaller tribes who sought to avoid Kikuyu domination by advocating a federal form of government. A coalition government of the two parties was formed in 1962, but after elections in May 1963 KANU took office, with Kenyatta as prime minister, under a new constitution that gave Kenya self-government. Following further discussions in London, Kenya became fully independent in December 1963, and Zanzibar surrendered its sovereignty over the Kenya Protectorate to Kenya. A year later Kenya became a republic, with Kenyatta as its first president and Oginga Odinga as vice president.

Formal independence

The Republic of Kenya. *Kenyatta's rule.* In 1964 an army mutiny was suppressed when Kenyatta sought the help of British troops. The president then acted decisively: better conditions of service were introduced, promotion prospects were improved, and the proportion of Kenyatta's own Kikuyu people in the officer corps was steadily increased. At the same time, most KADU members transferred their allegiance to KANU, and KADU ceased to exist. To forestall any new opposition, Kenyatta tried consistently to appoint members of different ethnic groups to official posts with all the patronage the appointments conferred, but he relied most heavily on Kikuyu, whom he believed he could trust. Ideological differences led to disagreement with Vice President Odinga, whose appointment had been made to satisfy the powerful Luo. Odinga believed that, by adopting a pro-Western, essentially capitalist economic policy, the government was neglecting the interests of poorer people. Breaking with KANU to form a new opposition party, the Kenya People's Union (KPU), Odinga found his position weakened by legislation requiring elected officials who switched parties to resign their seats and run for reelection. By contrast, Kenyatta's authority was strengthened because of the greater powers now attributed to the office of president.

The distribution to Africans of hundreds of thousands of smallholdings, made available by a satisfactory financial settlement with European farmers who were prepared to leave the country amicably, enabled Kenyatta to challenge Odinga's assertion that the poor were not catered to. The inability of some smallholders to offer the flexibility needed to meet the changing market demand for their produce meant that not all the new landowners were successful. There was, nevertheless, a sufficiently widespread improvement in living standards to ensure continuing support for the government.

The assassination in July 1969 of Mboya, who had become minister of economic planning and development, widened still further the serious ethnic gap that had already emerged as a result of Odinga's fall from grace. Mboya, like Odinga, was a Luo, though neither had pursued ethnic goals. Their supporters among the Luo, however, believed there was a Kikuyu plot, centring upon Kenyatta, which threatened Luo interests. The division grew greater after October 1969, when Odinga and some of his leading party members were arrested and the KPU was banned. The arrested men were subsequently released, and Odinga himself was freed in 1971 when the president endeavoured to achieve a measure of reconciliation and national unity. His efforts were not wholly successful, however, and the transfer of more than 1.5 million acres of land to a group of wealthy Kenyans, many of them Kikuyu, though defensible on grounds of potential economic efficiency, did not prove as profitable as had been hoped, largely because of bad management. Inevitably, the move provided fuel for those who, like Odinga, accused the government of being uncaring and for those who believed there was an ethnic plot to benefit the Kikuyu.

Suspicion of Kikuyu influence

A further suspected army coup attempt was forestalled in 1971. Once again Kenyatta responded by improving pay and conditions of service, but at the same time he made sure that the army remained small in numbers. Odinga and other former KPU leaders were prevented from standing in the parliamentary elections of 1974 by new regulations that forbade the candidacy of anyone who had not been a member of KANU for the past three years. The challenge to Kenyatta was then taken up in parliament by another former supporter of KANU, J.M. Kariuki. Kariuki was critical of growing corruption in the government, and he won considerable support when increasing oil prices and the consequent worldwide inflation caused hardship among the poorer members of the community. His arrest and murder in March 1975, which the government tried to cover up, aroused angry protests, particularly among university students whose criticisms of the government were to have increasing significance.

The question of who should succeed the aging president exacerbated the disagreements already prominent among the country's leaders. Kenyatta himself encouraged no one to claim the inheritance, but leading Kikuyu who had benefited greatly under his leadership plotted to secure a complaisant successor. The attorney general, Charles Njonjo, though himself a Kikuyu, opposed this move, as did another Kikuyu, the minister of finance, Mwai Kibaki. Together the two ensured that, upon Kenyatta's death in August 1978, he was succeeded by his deputy, Daniel arap Moi, a member of the minority Kalenjin group of peoples. Moi was elected president in October; Kibaki became vice president.

Moi's rule. Misgivings about what would happen on the departure of the dominating figure who had led his country to independence were soon dispelled. The transfer of power took place smoothly, owing mainly to the skillful leadership of Njonjo and Kibaki but also helped by a recent boom in coffee prices, which had eased the country's economic problems to a considerable extent. At first Moi followed Kenyatta's example by distributing offices among as wide an ethnic spectrum as possible, though over the

years members of his own Kalenjin group acquired a disproportionate number of appointments. Odinga remained critical of the government, and university students rallied to his support on idealistic grounds and also because they saw little prospect in the near future of being able to occupy the limited number of lucrative offices now held by their immediate predecessors.

Stifling opposition to the president

It was, however, Njonjo whom Moi saw as the main threat to his position. Exercising his almost unassailable presidential powers, he began a campaign of denigration against the former kingmaker. Njonjo's responsibilities were cut, as, too, were those of Kibaki. The strength of Moi's position was further underlined when the army loyally rallied to suppress an attempted coup by some of the lower ranks of the air force in 1982. Significantly, a number of the leaders of the rebellion were Luo, and many university students also took part in the disturbances. Anxious to prevent the students from becoming effective leaders of discontented groups of ethnic or impoverished minorities, the president resorted to closing the universities temporarily if opposition was voiced.

Generous financial support from the Western powers since independence had been an important factor in ensuring that Kenya's precarious economy survived the traumas of inflation. Kenyatta had set the pattern of aligning his country with the West, and Moi was happy to follow suit. The extension of land ownership among Africans in the postindependence period also helped to create a prosperous class that was anxious to avoid revolutionary change, while the powers vested in the office of president made successful opposition an unlikely prospect. (K.In.)

Increasingly, however, Western financial aid came to be tied to demands for political and economic reforms, and Moi finally accepted a 1991 constitutional amendment that reinstated multiparty elections. In elections held the following December, however, Moi was reelected, and the opposition was divided.

Moi made Njonjo chairman of the Kenya Wildlife Services and Richard Leakey head of the civil service and permanent secretary to the cabinet in 1999 (Leakey left in 2001). Leakey's popularity was cited as the main reason Moi appointed him to this post; it was also seen as Moi's way of showing Kenya's commitment to tackling the issues of corruption and government mismanagement, issues that continued to plague the country as did concerns over who would succeed Moi when he retired. Another major concern was the growing number of AIDS cases. (Ed.)

For later developments in the history of Kenya, see the BRITANNICA BOOK OF THE YEAR.

Tanzania

The United Republic of Tanzania (in Swahili, Jamhuri ya Muungano wa Tanzania) is situated just south of the Equator, bordered by the Indian Ocean on the east and eight other nations: Kenya, Uganda, Rwanda, Burundi, Congo (Kinshasa), Zambia, Malawi, and Mozambique. Tanzania was formed as a sovereign state in 1964 through the union of the theretofore separate states of Tanganyika and Zanzibar. The combined territories comprise 364,017 square miles (942,799 square kilometres), with mainland Tanganyika covering more than 99 percent of the total area. Mafia Island is administered from the mainland, while Zanzibar and Pemba islands have a separate government administration. Dodoma, since 1974 the official capital of Tanzania, is centrally located on the mainland. Dar es Salaam, however, remains the seat of most government administration and is the country's largest city.

PHYSICAL AND HUMAN GEOGRAPHY

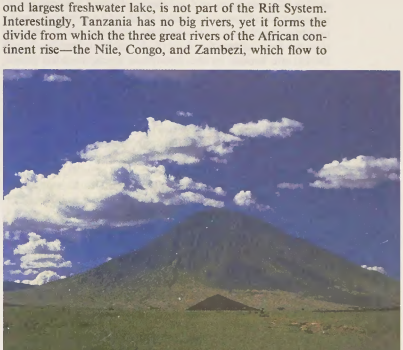
The land: Tanzania mainland. *Relief.* Except for the narrow coastal belt of the mainland and the offshore islands, most of Tanzania lies above 600 feet (200 metres) in elevation. Vast stretches of plains and plateaus contrast with spectacular relief features, notably Africa's highest mountain, Kilimanjaro (19,340 feet [5,895 metres]), and the world's second deepest lake, Lake Tanganyika (4,710 feet [1,436 metres] deep).

The East African Rift System runs in two north–south-trending branches through Tanzania, leaving many narrow, deep depressions that are often filled by lakes. One branch, the Western Rift Valley, runs along the western frontier and is marked by Lakes Tanganyika and Rukwa, while the other branch, the Eastern (or Great) Rift Valley, extends through central Tanzania from the Kenyan border in the region of Lakes Eyasi, Manyara, and Natron south to Lake Nyasa at the border with Mozambique. The Central Plateau, covering more than a third of the country, lies between the two branches.

The Rift Valley

Highlands associated with the Western Rift Valley are formed by the Ufipa Plateau, the Mbeya Range, and Rungwe Mountain in the southwestern corner of the country. From there the Southern Highlands run northeastward along the Great Rift to the Ukuguru and Nguru mountains northwest of Morogoro. Extending from the northern coast, the Usambara and Pare mountain chains run in a southeast-to-northwest direction, culminating in Kilimanjaro's lofty, snow-clad peak and continuing beyond to Mount Meru (14,980 feet). Immediately to the west of Mount Meru, another chain of mountains begins, which includes the still-active volcano Ol Doinyo Lengai and the Ngorongoro Crater, the world's largest caldera, or volcanic depression. This chain extends through a corridor between Lake Eyasi and Lake Manyara toward Dodoma.

Drainage. Because of its numerous lakes, approximately 22,800 square miles of Tanzania's territory consists of inland water. Lake Victoria, which ranks as the world's second largest freshwater lake, is not part of the Rift System. Interestingly, Tanzania has no big rivers, yet it forms the divide from which the three great rivers of the African continent rise—the Nile, Congo, and Zambezi, which flow to the Mediterranean Sea, the Atlantic Ocean, and the Indian Ocean, respectively. Separated by the Central Plateau, the watersheds of these rivers do not meet.



Ol Doinyo Lengai, volcano near Lake Natron, northern Tanzania
Robert Francis/Robert Harding Picture Library

All of Tanzania's major rivers—the Ruvuma, the Rufiji, the Wami, and the Pangani—drain into the Indian Ocean. The largest, the Rufiji River, has a drainage system that extends over most of southern Tanzania. The Kagera flows into Lake Victoria, whereas other minor rivers flow into internal basins formed by the Great Rift Valley. With so many rivers, Tanzania is rich in hydroelectric potential.

Soils. The variety of soils in Tanzania surpasses that of any other country in Africa. The reddish brown soils of volcanic origin in the highland areas are the most fertile. Many river basins also have fertile soils, but they are subject to flooding and require drainage control. The red and yellow tropical loams of the interior plateaus, on the

other hand, are of moderate-to-poor fertility. In these regions, high temperatures and low rainfall encourage rapid rates of oxidation, which result in a low humus content in the soil and, consequently, a clayey texture rather than the desired crumblike structure of temperate soils. Also, tropical downpours, often short in duration but very intense, compact the soil; this causes drainage problems and leaches the soil of nutrients.

Climate. Tanzania is subject to a warm, equatorial climate modified by variations in elevation. The high amount of solar radiation throughout the year is associated with a limited seasonal fluctuation of temperature: the mean monthly variation is less than 9° F (5° C) at most stations. Ground frosts rarely occur below 8,200 feet.

Rainfall and agriculture

Rainfall is highly seasonal, being influenced greatly by the annual migration of the intertropical convergence zone. Roughly half of Tanzania receives less than 30 inches (750 millimetres) of rainfall annually, an amount considered to be the minimum required for most forms of crop cultivation in the tropics. The Central Plateau, the driest area, with less than 20 inches per year on average, experiences a single rainy season between December and May. Rainfall is heavier on the coast, where there are two peaks of rainfall in October–November and April–May. The offshore islands and many highland areas have high annual rainfall totals of over 60 inches.

Plant and animal life. Forests grow in the highland areas where there is high rainfall and no marked dry season. The western and southern plateaus are primarily miombo woodland, consisting of an open cover of trees, notably *Brachystegia, Isoberlinia, Acacia,* and *Combretum.* In areas of less rainfall (16–32 inches), bushland and thicket are found. In the floodplain areas, wooded grassland with a canopy cover of less than 50 percent has been created by poor drainage and by the practice of burning for agriculture and animal grazing. Similarly, grassland appears where there is a lack of good drainage. For example, the famous Serengeti Plain owes its grasslands to a calcrete, or calcium-rich hardpan, deposited close to the surface by evaporated rainwater. Swamps are found in areas of perennial flooding. Desert and semidesert conditions range from an Alpine type at high altitudes to saline deserts in poorly drained areas and arid deserts in areas of extremely low rainfall.

Due to the historically low density of human settlement, Tanzania is home to an exceptionally rich array of wildlife. Large herds of hoofed animals—most spectacularly the wildebeest, as well as the zebra, giraffe, buffalo, gazelle, eland, dik-dik, and kudu—are found in most of the country's numerous game parks. Predators include hyenas, wild dogs, and the big cats—lions, leopards, and cheetahs. Crocodiles and hippopotamuses are common on riverbanks and lakeshores. The government has taken special measures to protect the rhinoceros and elephant, which have fallen victim to poachers. Small bands of chimpanzees inhabit Gombe National Park along Lake Tanganyika. Nearly 1,500 varieties of birds have been reported, and there are numerous species of snakes and lizards.

Settlement patterns. The two most important factors influencing the regional pattern of human settlement are rainfall and the incidence of tsetse fly. The tsetse, which thrives on wild game in miombo woodlands, is the carrier of *Trypanosoma,* a blood parasite that causes sleeping sickness in cattle and people. Tsetse infestation makes human settlement hazardous in areas of moderate rainfall, so that areas of low and unreliable rainfall are more densely populated than would otherwise be the case. The insect does not pose a threat to areas of high rainfall and high population density.

Centres of population

Population is concentrated in the highlands of the Mbeya Range, Kilimanjaro, and the Bukoba area west of Lake Victoria, on the cultivation steppe south of Lake Victoria, in the moderately high-rainfall region of Mtwara on the southern coast, and in the urban area of Dar es Salaam. These areas are all located on the perimeter of the country. The influence of the central rail line is clearly evident in the corridor of moderate population density extending from Dar es Salaam to Lake Victoria. Three sparsely populated areas stand out: the arid area of Arusha in the north and the two large tsetse-infested areas centred around Tabora in the west and Lindi and Songea in the south.

Regional variations in agricultural productivity are strongly related to the pressure of population on the land. Shifting cultivation, which involves rotating crops annually and leaving some fields to lie fallow for 20 years or more, was traditionally the means of renewing the fertility of the soil throughout Tanzania (with the exception of the more fertile and densely populated highland areas). A settlement pattern of widely dispersed, isolated farmsteads resulted from this practice. As the rural population expanded, however, fallow periods became shorter, and consequently soil fertility and crop yields suffered. In order to raise productivity, during the 1960s and '70s the government tried to bring about the use of improved agricultural methods, equipment, and fertilizer through the nucleation of rural settlements. First, the *ujamaa* (or "familyhood") policy of the 1960s supported collectivized agriculture in a number of government-sponsored planned settlements. These settlements were over-reliant on government finance and gradually dwindled in number. On a much larger scale, the "villagization" program of the 1970s moved millions of peasants into nucleated villages of 250 households or more. By 1978 there were more than 7,500 villages, in comparison to only about 800 in 1969. Villagization was aimed not at collectivizing agriculture but at facilitating the distribution of agricultural inputs such as fertilizers and improved seeds as well as making social services more accessible to the rural population.

Only about 15 percent of the population lives in urban areas, and one-third of the urban population resides in Dar es Salaam, a city of more than one million people. Bagamoyo and Tabora, old towns connected with the 19th-century Arab slave trade, have stagnated. The fortunes of Tanga, the second largest city during the British colonial period, have been tied to the export of sisal; as that has declined, the city has grown very slowly, although it remains the country's third largest city. Mwanza, Mbeya, and Arusha have thrived as trading centres remote from Dar es Salaam, and the growth of Morogoro and Moshi reflects their rich agricultural hinterlands. (D.F.Br.)

Important trading towns

The land: Zanzibar and Pemba. *Relief.* Low-lying Pemba, whose highest point reaches an elevation of 311 feet, and Zanzibar, which reaches 390 feet, are islands whose structure consists of coralline rocks. The west and northwest of Zanzibar consist of several ridges rising above 200 feet, but nearly two-thirds of the south and east are low-lying. Pemba appears hilly because the level central ridge has been gullied and eroded by streams draining into numerous creeks. On Zanzibar Island short streams drain mostly to the north and west. The few streams in the east disappear into the porous coralline rock.

Soils. Among the 10 types of soils recognized in Zanzibar are fertile sandy loams and deep red earths, which occur on high ground; on valley bottoms, less fertile gray and yellow sandy soils are found. The eight soil types in Pemba include brown loams; pockets of infertile sands are found on the plains.

Climate. Zanzibar and Pemba have rainfall of 60 inches and 80 inches, respectively. The rainfall is highest in April and May, lowest in November and December. Humidity is high. The average temperature is 81° F (27° C) in Zanzibar and 79° F (26° C) in Pemba; the annual temperature ranges are small.

Vegetation. Long human occupation has resulted in the clearance of most of the forests, which have been replaced with coconuts, cloves, bananas, citrus, and other crops. On the eastern side of the islands, especially on Zanzibar, there is bush (scrub).

Animal life. Although there is some difference between the animal life of the two islands, it is generally similar to that on the mainland. Animal life common to both islands includes monkeys, civet cats, and mongooses. More than 100 species of birds have been recorded in Zanzibar.

Rural settlements. The rural settlements—and the life in them—changed drastically after the nationalization of land in 1964 and subsequent agricultural reforms. By 1970, large plantations of cloves and coconuts, once al-

most exclusively the property of fewer than 50 Arab families, had been redistributed. The production of food crops, especially rice, is being encouraged. Fishing villages are still important in the east.

Urban settlements. The city of Zanzibar is still primarily a Muslim town, although the distinctive mode of life and culture, reminiscent of an Eastern commercial centre, has almost disappeared since the downfall of the Arab oligarchy in 1964. The hub of civic life is moving from Stone Town with its narrow lanes to a new town with modern buildings and amenities at Ngambo, the former African quarter. Kilimani, Bambi, and Chaani are being developed into new rural towns; a similar change is taking place in Pemba at Mkoani, Chake Chake, and Wete. (A.C.M.)

Ethnic variety

The people: Tanzania mainland. *Ethnic composition.* Tanzania is extremely heterogeneous, with more than 120 different indigenous African peoples as well as small groups of Asians and Europeans. As early as 5000 BC, Santype hunting bands inhabited the country. The Sandawe hunters of northern Tanzania are thought to be their descendants. By 1000 BC, agriculture and pastoral practices were being introduced through the migration of Cushitic people from Ethiopia. The Iraqw, Mbugu, Gorowa, and Burungi have Cushitic origins. About AD 500, iron-using Bantu agriculturalists coming from the west and south started displacing or absorbing the San hunters and gatherers; at roughly the same time, Nilotic pastoralists entered the area from the southern Sudan. Today the majority of Tanzanians are of Bantu descent; the Sukuma constitute the largest group, and others are the Nyamwezi, Hehe, Nyakyusa, Makonde, Yao, Haya, Chaga, Gogo, and Ha. Nilotic peoples are represented by the Masai, Arusha, Samburu, and Baraguyu. No one group has been politically or culturally dominant, although the tribes that were subject to Christian missionary influence and Western education during the colonial period (notably the Chaga and Haya) are now disproportionately represented in the government administration and cash economy.

There are also Asian and European minorities. During the colonial period, Asian immigration was encouraged, and Asians dominated the up-country produce trade. Coming mostly from Gujurāt in India, they form several groups distinguished by religious belief: the Ismā'ilis, Bohrās, Sikhs, Punjabis, and Goans. Since independence the Asian population has steadily declined due to emigration. The European population, never large because Tanganyika was not a settler colony, was made up primarily of English, Germans, and Greeks. In the postindependence period, a proliferation of different European, North American, and Japanese expatriates connected with foreign aid projects have made Tanzania their temporary residence.

Language. Swahili is the national language. Virtually all Tanzanians speak the language, and it is used as the medium of instruction in the first seven years of primary education. English, the country's second official language (together with Swahili), is the medium of instruction at further levels of education and is commonly used by the government in official business. Most African Tanzanians speak their traditional tribal language as well. The main languages spoken by the Asian minorities are Gujarati, Hindi, Punjabi, and Urdu.

Religion. Roughly one-third of the population is Muslim, another third professes Christianity, and the remainder is considered to hold animist beliefs. The division is usually not as clear as official statistics suggest, since many rural Tanzanians adhere to elements of their traditional animistic religions while practicing their Islāmic or Christian faith. A wide array of Christianity is represented, notably Lutheranism and Roman Catholicism. Among Muslims, both the Sunnite and Shi'ite sects are represented. The majority of Asian Muslims are Ismā'ili Khōjās under the Aga Khan's spiritual leadership. In addition, there are Asian adherents to Hinduism, Jainism, and Roman Catholicism.

The growing population

Demographic trends. Tanzania's population growth rate is one of the highest in sub-Saharan Africa. Despite great improvements since the 1950s, the infant and child mortality rates remain high, but fertility is high as well. Nearly 50 percent of Tanzanians are under the age of 15. Life expectancy, at about 53 years, is above average for the subcontinent. (D.F.Br.)

The people: Zanzibar and Pemba. *Language.* Swahili is the principal language in Zanzibar and Pemba. The classical dialect is Kiunguja. Arabic is also important because of long-established Islāmic tradition, past Arab overlordship, and the presence of a large Arab-speaking minority. Among the Asian communities, the chief languages are Gujarati, Kutchi, and Hindustani. English, taught in schools, is widely used.

Ethnic groups. There are several groups of Africans, nearly one-third of whom are recent arrivals from the mainland. Indigenous Bantu groups, consisting of the Pemba in Pemba and the Hadimu and Tumbatu in Zanzibar, have absorbed the settlers who came from Persia in the 10th century. These groups and some of the descendants of slaves call themselves Shirazi. There are also small enclaves of Comorians and Somalis. Arab settlements were also established early, and intermarriage with the local people took place. Later Arab arrivals came from Oman and constituted an elite. The poorer recent immigrants from Oman are known as Manga. The Asians, now forming a very small minority, may be divided into Muslim and non-Muslim groups.

Religion. Almost the whole of the Arab and the African peoples of Zanzibar profess the Islāmic faith. Traditional African beliefs are also found existing in conjunction with Islām. Among Muslims, the Sunnite sect is preferred by the indigenous people. (A.C.M.)

The economy. The Tanzanian economy is overwhelmingly agrarian in nature and reflects the leadership's political commitment to socialist development and central planning. Agriculture constitutes over half of the gross domestic product (GDP) and some 80 percent of export earnings, and it provides a livelihood for about nine-tenths of the economically active population. Industry accounts for less than 10 percent of the GDP, and mining less than 1 percent, whereas services, including public administration, produce approximately one-third of the GDP. A number of industries and public services were nationalized at the time of the Arusha Declaration in 1967, when the intention to build a socialist state was announced.

Economic troubles

Beginning in 1979 and continuing throughout the 1980s, the international oil price rise, the country's declining terms of trade, and the sluggishness of the domestic economy brought about rapid inflation, the emergence of an unofficial market (consisting of the smuggling of goods abroad in order to avoid taxes and price controls), and a government fiscal crisis. Despite attempts to cut imports to the barest minimum, the trade deficit widened to an unprecedented level, and the balance-of-payments problem became so acute that development projects had to be suspended. This economic crisis forced the government to secure a loan from the International Monetary Fund (IMF) in 1986. The loan's conditions required the elimination of subsidies and price controls as well as some social services and staff positions in state-run enterprises. Thereafter, the government continued to implement measures intended to create a mixed economy and reduce the extent of the untaxable unofficial markets.

Agriculture. The major food crops are corn, rice, sorghum, millet, bananas, cassava, sweet potatoes, barley, potatoes, and wheat. Corn and rice are the preferred cereals, whereas cassava and sweet potatoes are used as famine-prevention crops owing to their drought-resistant qualities. In some areas food crops are sold as cash crops. Peasants in the Ruvuma and Rukwa regions, for example, have specialized in commercial corn production, and in riverine areas, especially along the Rufiji, rice is sold.

Export cash crops provide the major source of foreign exchange for the country. Coffee and cotton are by far the most important in this respect, but exports of tea, cashew nuts, tobacco, and sisal are also substantial. Cloves are Zanzibar's main export. Once the source of over 90 percent of the world's cloves, Zanzibar now produces only about 10 percent of the international supply.

The villagization program of the mid-1970s was followed by government efforts to distribute improved seed corn and fertilizers through the new village administrations, but

timely distribution of such agricultural inputs was largely thwarted by the logistical problems of transporting them to the villages. Nevertheless, increased yields, attributed to the use of chemical fertilizers, have been achieved in peasant corn production in the south and southwestern regions.

Industry. Tanzania's industry is based on the processing of its agricultural goods and on import substitution—that is, the manufacture (often from imported materials and parts) of products that were once purchased from abroad. The principal industries are food processing, textiles, brewing, and cigarettes. Production of clothing, footwear, tires, batteries, and bottles takes place as well. There is also a hot-rolling steel mill, a factory producing asbestos cement sheets, a bicycle factory, and a large pulp and paper mill.

A strategy to lay the foundation for the rapid growth of such basic industries as steel, chemicals, rubber, and textiles was thwarted by the national economic crisis of the 1980s. The large amounts of imported materials, parts, and capital equipment necessary to implement such a policy could not be paid for, owing to the country's lack of foreign exchange. Standby credit facilities from the IMF provided the capital investment needed to initiate a rehabilitation of industry.

Mineral deposits

Resources. Tanzania mines diamonds, gold, kaolin, gypsum, tin, and various gemstones, including tanzanite. There are large exploitable deposits of coal in the southwest, phosphate deposits in Arusha, and nickel in the Kagera region. Natural gas has been discovered at Songo Songo Island. Several international companies have been involved in onshore and offshore petroleum exploration.

Tanzania's native forests are primarily composed of hardwoods, but softwood production is increasing. A large pulp and paper mill at Mufindi is supplied by the extensive softwood forest nearby at Sao Hill. Because firewood and charcoal are the major domestic fuels, there is growing concern about the deforestation of land in the hinterland of Dar es Salaam and Tanzania's other large towns.

Several lakes, especially Lake Victoria, are important sources of fish. Prawns are commercially fished in the Rufiji River delta, but coastal fishery is primarily of an artisanal nature.

Finance and trade. All private banks were nationalized between 1967 and 1992, but since then private banks (including branches of foreign-owned banks) have been allowed to open. The state-run Bank of Tanzania operates as the central bank; it manages the country's finances and issues the currency, the Tanzanian shilling.

Transportation. Transport in Tanzania spans a wide spectrum, from the motorized means made possible by roads, railways, seaports, and airfields to the traditional carrying of loads by animals and people. It is estimated that the average peasant household carries by headload approximately 56 ton-miles (90 ton-kilometres) of firewood, water, and crops per year.

Roads are by far the most important nontraditional mode of transport, carrying some 70 percent of total traffic. The

The road network

road network extends to all parts of the country, but it is densest along the coast and southeast of Lake Victoria. The country's percentage of roads paved is one of the lowest in sub-Saharan Africa. The Tanzam Highway, opened in the early 1970s between Dar es Salaam and Zambia, has significantly reduced the isolation of southern Tanzania. A newer highway intersects it at Makambako and proceeds southward through the southern highlands to Songea. Government efforts have been placed on rehabilitating the trunk road system, which deteriorated with a decline in the importation of maintenance materials during the economic crisis.

Dar es Salaam port, with its deep-water berths, handles about three-fourths of all ships calling at Tanzanian ports. The remainder go primarily to Tanga, Mtwara, and the port of the city of Zanzibar. The Tanzania Coastal Shipping Line offers transport services along the coast; a passenger ferry operates between Dar es Salaam and Zanzibar.

The Air Tanzania Corporation provides internal air services as well as international flights to destinations in central and southern Africa, the Persian Gulf, and the Indian Ocean. There are international airports at Dar es Salaam, Kilimanjaro, and Zanzibar, but most scheduled international flights land in Dar es Salaam.

The railway system dates back to the pre-World War I, German-built Central Railway Line, which bisects the country between Dar es Salaam and Kigoma, and the Tanga-to-Moshi railway. Today there is also a branch between these two lines, and another line connects Mwanza with Tabora on the Central Line. The TAZARA rail line, running between Dar es Salaam and Kapiri-Mposhi on the Zambian border, was built with Chinese aid in the early 1970s.

Administration and social conditions. *The government.* The Interim Constitution of 1965 established the United Republic of Tanzania through the merger of Tanganyika and Zanzibar, until then separate and independent countries. A permanent constitution for the United Republic was approved in 1977 and amended in 1984 to include a bill of rights.

Zanzibar has a separate constitution, approved in 1979 and amended in 1985. The executive branch is composed of an elected president and a cabinet called the Supreme Revolutionary Council. Zanzibar's parliament, the House of Representatives, is made up of elected and appointed members. These political bodies deal with matters internal to Zanzibar. Since the union with Tanganyika, some segments of Zanzibari society have occasionally demanded greater autonomy from the mainland. (D.F.Br.)

The president of the United Republic is the head of state and commander in chief of the armed forces. The cabinet of ministers is advisory to the president. Prior to 1995 it included two vice presidents: the prime minister, who is appointed by the president and acts as the leader of the cabinet, and the president of Zanzibar. Since then an amendment to the constitution, which was approved in 1994 and took effect after the 1995 general election, rescinded the stipulation that called for the president of Zanzibar to serve as a vice president.

According to the 1984 constitutional amendments, most members (216 in the 1990 election) of the National Assembly are directly elected. Many seats also are allocated to ex-officio, nominated, and indirectly elected members—including those seats reserved for women, representatives of mass organizations, and the president's nominees. The National Assembly has a term of five years but can be dissolved by the president before this term expires.

The National Assembly

By law Tanzania was a one-party state until 1992, when the constitution was amended to establish a multiparty political process. In 1977 the Tanganyika African Nationalist Union, which had led the colony to independence, and the Afro-Shirazi Party of Zanzibar, which had taken power after a coup in 1964, were amalgamated into the Revolutionary Party (Chama cha Mapinduzi; CCM). Prior to the 1992 amendment, the CCM dominated all aspects of political life, and there was no clear separation of party and government personnel at regional and district levels. By the time of the first national multiparty election in 1995, more than a dozen opposition political movements were officially registered.

For administrative purposes, mainland Tanzania is divided into 20 regions. Each region is administered by a commissioner who is appointed by the central government. At district, division, and ward levels, there are popularly elected councils with appointed executive officers. (D.F.Br./Ed.)

The judiciary. Tanzania's judiciary is appointed by the president in consultation with the chief justice. Judges cannot be dismissed except on the grounds of misbehaviour or incapacity. A network of primary and district courts has been established throughout the country. English, Islāmic, and customary laws have been absorbed into the legal system. In Zanzibar the highest judicial authority is the Supreme Council. Muslim courts deal with marriage, divorce, and inheritance.

Education. The government-supported education system has three levels: primary (seven years), secondary (four to six years), and university, as well as vocational training schools. During the mid-1970s universal primary education was made mandatory, resulting in a vast increase in primary-school children. Popular pressure for

the expansion of secondary schools has outstripped the availability of government finance. As a result, private secondary schools sponsored by religious institutions and, most notably, by parents themselves have expanded in number. There are two universities, the University of Dar es Salaam (1961), formerly part of the University of East Africa, and Sokoine University of Agriculture (1984). Extensive adult education has focused on eradicating illiteracy, and, as a result, Tanzania has one of the highest literacy rates in Africa.

Health and welfare. National and local governments support a network of village dispensaries and rural health centres; hospitals are located in the urban areas. Private doctors and religious organizations provide medical facilities as well.

Preventing disease

The emphasis of national health policy has been on preventive medicine, especially better nutrition, maternal and child health, environmental sanitation, and the prevention and control of communicable diseases. The main communicable diseases are poliomyelitis, leprosy, tuberculosis, dysentery, enteric fevers, and AIDS. Environmental diseases include malaria, sleeping sickness, bilharzia, and onchocerciasis (river blindness). Inadequate nutrition, particularly of children, is a major concern. Improvements in health and reduction of mortality rates have resulted from the provision of medical care to the rural population and from an inoculation program for children.

Cultural life. Olduvai Gorge, in the Great Rift Valley, is the site of the discovery of some of the earliest known remains of human ancestry, dating back 1,750,000 years. The ancient in-migration of Cushitic, Nilotic, and Bantu peoples, displacing the native San-type population, resulted in a complex agglomeration of tribal communities practicing complementary forms of pastoral and agricultural livelihoods. In the last 500 years, Portuguese, Arab, Indian, German, and British traders and colonists have added to the mosaic. Today Tanzania's multiethnic and multiracial population practices a variety of traditions and customs that form a rich cultural heritage.

The role of kin is central to Tanzanian social and recreational life. Visiting kin on joyous and sorrowful family occasions is given high priority despite the inconvenience caused by a relatively undeveloped transport system. Educated members of the extended family are frequently held responsible for the education and welfare of younger siblings.

Association football (soccer) is a popular sport. In international competitions, Tanzanian sportsmen have excelled in long-distance running.

Oral storytelling traditions and tribal dancing are an important part of the cultural life of the rural population. The University of Dar es Salaam has an active theatre arts group. Among the visual arts, Makonde carvers from southern Tanzania are renowned for their abstract ebony carvings, and Zanzibar is famous for its elaborately carved doors and Arab chests. Basket weaving, pottery, and musical instrument making are prevalent in many rural areas.

Tanzania has government-owned Swahili and English daily newspapers. The radio, more than newspapers or television, is the medium through which the rural population receives national and international news. The radio has been extensively used by the government for the promotion of adult literacy, better nutrition, and ecological conservation. (D.F.Br.)

For statistical data on the land and people of Tanzania, see the *Britannica World Data* section in the BRITANNICA BOOK OF THE YEAR.

HISTORY

Tanganyika. *Early exploration.* Most of the known history of Tanganyika before the 19th century concerns the coastal area, although the interior has a number of important prehistoric sites, including the Olduvai Gorge. Trading contacts between Arabia and the East African coast existed by the 1st century AD, and there are indications of connections with India. The coastal trading centres were mainly Arab settlements, and relations between the Arabs and their African neighbours appear to have been fairly friendly. After the arrival of the Portuguese in the late 15th century, the position of the Arabs was gradually undermined, but the Portuguese made little attempt to penetrate into the interior. They lost their foothold north of the Ruvuma River early in the 18th century as a result of an alliance between the coastal Arabs and the ruler of Muscat on the Arabian Peninsula. This link remained extremely tenuous, however, until French interest in the slave trade from the ancient town of Kilwa, on the Tanganyikan coast, revived the trade in 1776. Attention by the French also aroused the sultan of Muscat's interest in the economic possibilities of the East African coast, and a new Omani governor was appointed at Kilwa. For some time most of the slaves came from the Kilwa hinterland, and until the 19th century such contacts as existed between the coast and the interior were due mainly to African caravans from the interior.

Early trade on the coast

In their constant search for slaves, Arab traders began to penetrate farther into the interior, more particularly in the southeast toward Lake Nyasa. Farther north two merchants from India followed the tribal trade routes to reach the country of the Nyamwezi about 1825. Along this route ivory appears to have been as great an attraction as slaves, and Sa'id ibn Sultan himself, after the transfer of his capital from Muscat to Zanzibar, gave every encouragement to the Arabs to pursue these trading possibilities. From the Nyamwezi country the Arabs pressed on to Lake Tanganyika in the early 1840s. Tabora (or Kazé, as it was then called) and Ujiji, on Lake Tanganyika, became important trading centres, and a number of Arabs made their homes there. They did not annex these territories but occasionally ejected hostile chieftains. Mirambo, an African chief who built for himself a temporary empire to the west of Tabora in the 1860s and '70s, effectively blocked the Arab trade routes when they refused to pay him tribute. His empire was purely a personal one, however, and collapsed on his death in 1884.

The first Europeans to show an interest in Tanganyika in the 19th century were missionaries of the Church Missionary Society, Johann Ludwig Krapf and Johannes Rebmann, who in the late 1840s reached Kilimanjaro. It was a fellow missionary, Jakob Erhardt, whose famous "slug" map (showing, on Arab information, a vast, shapeless, inland lake) helped stimulate the interest of the British explorers Richard Burton and John Hanning Speke. They traveled from Bagamoyo to Lake Tanganyika in 1857–58, and Speke also saw Lake Victoria. This expedition was followed by Speke's second journey, in 1860, in the company of J.A. Grant, to justify the former's claim that the Nile rose in Lake Victoria. These primarily geographic explorations were followed by the activities of David Livingstone, who in 1866 set out on his last journey for Lake Nyasa. Livingstone's object was to expose the horrors of the slave trade and, by opening up legitimate trade with the interior, to destroy the slave trade at its roots. Livingstone's journey led to the later expeditions of H.M. Stanley and V.L. Cameron. Spurred on by Livingstone's work and example, a number of missionary societies began to take an interest in East Africa after 1860.

David Livingstone

German East Africa. It was left to Germany, with its newly awakened interest in colonial expansion, to open up the country to European influences. The first agent of German imperialism was Carl Peters, who, with Joachim, Count Pfeil and Karl Juhlke, evaded the sultan of Zanzibar late in 1884 to land on the mainland. He made a number of "contracts" in the Usambara area by which several chiefs were said to have surrendered their territory to him. Peters' activities were confirmed by Bismarck. By the Anglo-German Agreement of 1886 the sultan of Zanzibar's vaguely substantiated claims to dominion on the mainland were limited to a 10-mile-wide coastal strip, and Britain and Germany divided the hinterland between them as spheres of influence, the region to the south becoming known as German East Africa. Following the example of the British to the north, the Germans obtained a lease of the coastal strip from the sultan in 1888, but their tactlessness and fear of commercial competition led to a Muslim rising in August 1888. The rebellion was put down only after the intervention of the imperial German government and with the assistance of the British navy.

Recognizing the administrative inability of the German East Africa Company, which had theretofore ruled the country, the German government declared a protectorate over its sphere of influence in 1891 and over the coastal strip, where the company had bought out the sultan's rights. Germany was anxious to exploit the resources of its new dependency, but lack of communications at first restricted development to the coastal area. The introduction of sisal from Florida in 1892 by the German agronomist Richard Hindorff marked the beginning of the territory's most valuable industry, which was encouraged by the development of a railway from the new capital of Dar es Salaam to Lake Tanganyika. In 1896 work began on the construction of a railway running northeastward from Tanga to Moshi, which it reached in 1912. This successfully encouraged the pioneer coffee-growing activities on the slopes of Kilimanjaro. Wild rubber tapped by Africans, together with plantation-grown rubber, helped swell the country's economy. The government also supplied good-quality cottonseed free to African growers and sold it cheaply to European planters. The administration tried to make good the lack of clerks and minor craftsmen by encouraging the development of schools, an activity in which various missionary societies were already engaged.

Resistance to German rule

The enforcement of German overlordship was strongly resisted, but control was established by the beginning of the 20th century. Almost at once came a reaction to German methods of administration, the outbreak of the Maji Maji rising in 1905. Although there was little organization behind it, the rising spread over a considerable portion of southeastern Tanganyika and was not finally suppressed until 1907. It led to a reappraisal of German policy in East Africa. The imperial government had attempted to protect African land rights in 1895 but had failed in its objective in the Kilimanjaro area. Similarly, liberal labour legislation had not been properly implemented. The German government set up a separate Colonial Department in 1907, and more money was invested in East Africa. A more liberal form of administration rapidly replaced the previous semimilitary system.

World War I put an end to all German experiments. Blockaded by the British navy, the country could neither export produce nor get help from Germany. The British advance into German territory continued steadily from 1916 until the whole country was eventually occupied. The effects of the war upon Germany's achievements in East Africa were disastrous; the administration and economy were completely disrupted. In these circumstances the Africans reverted to their old social systems and their old form of subsistence farming. Under the Treaty of Versailles (1919), Britain received a League of Nations mandate to administer the territory except for Ruanda-Urundi, which came under Belgian administration, and the Kionga triangle, which went to Portugal.

Tanganyika Territory. Sir Horace Byatt, administrator of the captured territory and, from 1920 to 1924, first British governor and commander in chief of Tanganyika Territory (as it was then renamed), enforced a period of recuperation before new development plans were set on foot. A Land Ordinance (1923) ensured that African land rights were secure. Sir Donald Cameron, governor from 1925 to 1931, infused a new vigour into the country. He reorganized the system of native administration by the Native Authority Ordinance (1926) and the Native Courts Ordinance (1929). His object was to build up local government on the basis of traditional authorities, an aim that he pursued with doctrinaire enthusiasm and success. He attempted to silence the criticisms by Europeans that had been leveled against his predecessor by urging the creation of a Legislative Council in 1926 with a reasonable number of nonofficial members, both European and Asian. In his campaign to develop the country's economy, Cameron won a victory over opposition from Kenya by gaining the British government's approval for an extension of the Central Railway Line from Tabora to Mwanza (1928). His attitude toward European settlers was determined by their potential contribution to the country's economy. He was, therefore, surprised by the British government's reluctance to permit settlement in Tanganyika. The economic depression after 1929 resulted in the curtailment of many of Cameron's development proposals. In the 1930s, too, Tanganyika was hampered by fears that it might be handed back to Germany in response to Hitler's demands for overseas possessions.

Economic development

At the outbreak of World War II Tanganyika's main task was to make itself as independent as possible of imported goods. Inevitably the retrenchment evident in the 1930s became still more severe, and, while prices for primary products soared, the value of money depreciated proportionately. Tanganyika's main objective after the war was to ensure that its program for economic recovery and development should go ahead. The continuing demand for primary produce strengthened the country's financial position. The chief item in the development program was a plan to devote 3 million acres (1.2 million hectares) of land to the production of peanuts (the Groundnuts Scheme). The plan, which was to be financed by the British government, was to cost £25 million, and, in addition, a further £4.5 million would be required for the construction of a railway in southern Tanganyika. It failed because of the lack of adequate preliminary investigations and was subsequently carried out on a greatly reduced scale.

Constitutionally, the most important immediate postwar development was the British government's decision to place Tanganyika under UN trusteeship (1947). Under the terms of the trusteeship agreement, Britain was called upon to develop the political life of the territory, which, however, only gradually began to take shape in the 1950s with the growth of the Tanganyika African National Union (TANU). The first two African members had been nominated to the Legislative Council in December 1945. This number was subsequently increased to four, with three Asian nonofficial members and four Europeans. An official majority was retained. In an important advance in 1955, the three races were given parity of representation on the unofficial side of the council with 10 nominated members each, and for a time it seemed as if this basis would persist. The first elections to the unofficial side of the council, however, enabled TANU to show its strength, for even among the European and Asian candidates only those supported by TANU were elected.

A constitutional committee in 1959 unanimously recommended that after the elections in 1960 a large majority of the members of both sides of the council should be Africans and that elected members should form the basis of the government. The approval of the British colonial secretary was obtained for these proposals in December 1959, and in September 1960 a predominantly TANU government took office. The emergence of this party and its triumph over the political apathy of the people were largely due to the leadership of Julius Nyerere. Tanganyika became independent on Dec. 9, 1961, with Nyerere as its first prime minister.

Zanzibar. The history of Zanzibar has been to a large extent shaped by the monsoons (prevailing trade winds) and by its proximity to the continent. The regular annual recurrence of the monsoons has made possible its close connection with India and the countries bordering the Red Sea and the Persian Gulf. Its proximity to the continent has made it a suitable jumping-off point for trading and exploring ventures not only along the coast but also into the interior.

Contact with Islāmic states

Portuguese and Omani domination. Though the first references to Zanzibar occur only after the rise of Islām, there would appear to be little doubt that its close connection with southern Arabia and the countries bordering the Persian Gulf began before the Christian era. At the beginning of the 13th century, the Arab geographer Yakut recorded that the people of Lenguja (namely, Unguja, the Swahili name for Zanzibar) had taken refuge from their enemies on Tumbatu, the inhabitants of which were Muslims.

In 1498 Vasco da Gama visited Malindi, and in 1503 Zanzibar Island was attacked and made tributary by the Portuguese. It appears to have remained in that condition for about a quarter of a century. Thereafter the relations between the rulers of Zanzibar and the Portuguese seem to have been those of allies, the people of Zanzibar more

than once cooperating with the Portuguese in attacks upon Mombasa. In 1571 the "king" of Zanzibar, in gratitude for Portuguese assistance in expelling certain African invaders, donated the island to his allies, but the donation was never implemented. A Portuguese trading factory and an Augustinian mission were established on the site of the modern city of Zanzibar, and a few Portuguese appear also to have settled as farmers in different parts of the island. The first English ship to visit Zanzibar (1591–92) was the *Edward Bonaventure,* captained by Sir James Lancaster.

When the Arabs captured Mombasa in 1698, all these settlements were abandoned and (except for a brief Portuguese reoccupation in 1728) Zanzibar and Pemba came under the domination of the Arab rulers of Oman. For more than a century those rulers left the government of Zanzibar to local hakims (governors). The first sultan to take up residence in Zanzibar was Sayyid Sa'īd ibn Sulṭān, who after several short visits settled there soon after 1830 and subsequently greatly extended his influence along the East African coast. On Sa'īd's death in 1856 his son Majīd succeeded to his African dominions, while another son, Thuwayn, succeeded to Oman.

As a result of an award made in 1860 by Lord Canning, governor general of India, the former African dominions of Sa'īd were declared to be independent of Oman. Majīd died in 1870 and was succeeded by his brother Barghash. Toward the end of the latter's reign his claims to dominion on the mainland were restricted by Britain, France, and Germany to a 10-mile-wide coastal strip, the administration of which was subsequently shared by Germany and Britain. Barghash died in 1888. Both he and Majīd had acted largely under the influence of Sir John Kirk, who was British consular representative at Zanzibar from 1866 to 1887. It was by Kirk's efforts that Barghash consented in 1873 to a treaty for the suppression of the slave trade.

British protectorate. In 1890 what was left of the sultanate was proclaimed a British protectorate, and in 1891 a constitutional government was instituted under British auspices, with Sir Lloyd Mathews as first minister. In August 1896, on the death of the ruling sultan, Ḥamad ibn Thuwayn, the royal palace at Zanzibar was seized by Khālid, a son of Sultan Barghash, who proclaimed himself sultan. The British government disapproved, and, as he refused to submit, the palace was bombarded by British warships. Khālid escaped and took refuge at the German consulate, whence he was conveyed to German East Africa. Ḥamud ibn Muḥammed was then installed as sultan (Aug. 27, 1896). In 1897 the legal status of slavery was finally abolished. In 1913 the control of the protectorate passed from the Foreign Office to the Colonial Office, when the posts of consul general and first minister were merged into that of British resident. At the same time, a Protectorate Council was constituted as an advisory body. In 1926 the advisory council was replaced by nominated executive and legislative councils.

Enforcement of British rule

Khalīfa ibn Harūb became sultan in 1911. He was the leading Muslim prince in East Africa, and his moderating influence did much to steady Muslim opinion in that part of Africa at times of political crisis, especially during the two world wars. He died on Oct. 9, 1960, and was succeeded by his eldest son, Sir Abdullah ibn Khalīfa.

In November 1960 the British Parliament approved a new constitution for Zanzibar. The first elections to the Legislative Council then established were held in January 1961 and ended in a deadlock. Further elections, held in June, were marked by serious rioting and heavy casualties. Ten seats were won by the Afro-Shirazi Party (ASP), representing mainly the African population; 10 by the Zanzibar Nationalist Party (ZNP), representing mainly the Zanzibari Arabs; and 3 by the Zanzibar and Pemba People's Party (ZPPP), an offshoot of the ZNP. The ZNP and ZPPP combined to form a government with Mohammed Shamte Hamadi as chief minister.

A constitutional conference held in London in 1962 was unable to fix a date for the introduction of internal self-government or for independence, because of failure to agree on franchise qualifications, the number of elected seats in the legislature, and the timing of the elections. An independent commission, however, subsequently delimited new constituencies and recommended an increase in the numbers of the Legislative Council, which the council accepted, also agreeing to the introduction of universal adult suffrage. Internal self-government was established in June 1963, and elections held the following month resulted in a victory for the ZNP–ZPPP coalition, which won 18 seats, the ASP winning the remaining 13. Final arrangements for independence were made at a conference in London in September. In October it was agreed that the Kenya coastal strip—a territory that extended 10 miles inland along the Kenya coast from the Tanganyika frontier to Kipini and that had long been administered by Kenya although nominally under the sovereignty of Zanzibar—would become an integral part of Kenya on that country's attainment of independence.

Independence. On Dec. 10, 1963, Zanzibar achieved independence as a member of the Commonwealth. In January 1964 the Zanzibar government was overthrown by an internal revolution, Sayyid Jamshid ibn Abdullah (who had succeeded to the sultanate in July 1963 on his father's death) was deposed, and a republic was proclaimed.

Revolt against Arab rule

Although the revolution was carried out by only about 600 armed men under the leadership of the communist-trained "field marshal" John Okello, it won considerable support from the African population. Thousands of Arabs were massacred in riots, and thousands more fled the island. Sheikh Abeid Amani Karume, leader of the Afro-Shirazi Party, was installed as president of the People's Republic of Zanzibar and Pemba. Sheikh Abdulla Kassim Hanga was appointed prime minister, and Abdul Raḥman Mohammed ("Babu"), leader of the new left-wing Umma (The Masses) Party (formed by defectors from the ZNP), became minister for defense and external affairs. Pending the establishment of a new constitution, the cabinet and all government departments were placed under the control of a Revolutionary Council of 30 members, which was also vested with temporary legislative powers. Zanzibar was proclaimed a one-party state. Measures taken by the new government included the nationalization of all land, with further powers to confiscate any immovable property without compensation except in cases of undue hardship.

The United Republic. The Tanganyikan constitution was amended in 1962, and Julius Nyerere became executive president of the Republic of Tanganyika. In 1963 TANU was declared the only legal party, but voters in each constituency were often offered a choice between more than one TANU candidate in parliamentary elections. That this arrangement amounted to something more than lip service to the idea of democracy was demonstrated in 1965 and in subsequent elections when, although Nyerere was reelected again and again as the sole candidate for president, a considerable number of legislators, including cabinet ministers, lost their seats.

An army mutiny was suppressed in January 1964 only after the president had reluctantly sought the assistance of British marines. Nyerere's authority was quickly restored, however, and in April he made an agreement with President Karume of Zanzibar to establish the United Republic of Tanzania, with Nyerere himself as president and Karume as first vice president. (Despite unification, for years Zanzibar continued to pursue its own policies, paying little attention to mainland practices.)

Nyerere's chief external task was to convince the outside world, particularly the Western powers, that Tanzania's foreign policy was to be one of nonalignment; but the overt involvement of the Eastern bloc in Zanzibar, as well as Nyerere's own insistence that to rectify the imbalance created in the colonial era Tanzania must turn more to the East for aid, did little to make the task easier. The high moral tone taken by the president over Britain's role in Rhodesia and over the supply of British arms to South Africa also strained the bonds of friendship between the two countries, with Tanzania severing diplomatic relations with Britain from 1965 to 1968. The consequent loss of aid from Britain was more than made good by help from Eastern countries, notably from China, which culminated in 1970 in the offer of an interest-free Chinese loan to finance the construction of a railway line linking Dar es Salaam with Zambia.

Though Nyerere fully appreciated the generous assistance his country was receiving, he was anxious to impress upon his countrymen the need for maximum self-reliance. Political freedom, he insisted, was useless if the country was to be enslaved by foreign investors. His views were formulated in the Arusha Declaration of Feb. 5, 1967. The resources of the country, Nyerere said, were owned by the whole people and were held in trust for their descendants. The leaders must set an example by rejecting the perquisites of a capitalist system and should draw only one salary. Banks must be nationalized, though compensation would be given to shareholders; the same would apply to the more important commercial companies. Agriculture, however, was the key to development, and only greater productivity could hold at bay the spectre of poverty. To give a fillip to his argument, people were to be moved into cooperative villages where they could work together for their mutual benefit.

Nyerere's Arusha Declaration

Nyerere's exhortations did not arouse the enthusiasm for which he had hoped. Individuals resisted his plans for collectivization, and not even the majority of his supporters wholeheartedly adopted his moral stand. The cooperative village scheme failed, bringing additional pressure to bear upon an already desperately weak economy. The sisal industry, one of those nationalized, became badly run down by the mid-1970s because of inefficient management.

Nyerere's criticisms were not reserved for his own people, nor yet for the wealthy nations of the world. In 1968 he challenged the rules of the Organization of African Unity (OAU) by recognizing the secession of Biafra from Nigeria, and in 1975 he attacked the OAU for planning to hold its summit meeting in Uganda, where Idi Amin was acting with extreme cruelty. Relations with Kenya also deteriorated, and in 1977 the East African Community ceased to exist after the closure of the Kenyan border. Even more challenging to the OAU's policy of nonintervention in the affairs of member states was the attack launched against Uganda in January 1979 after Amin had invaded the northwestern corner of Tanzania in October 1978. The retention of Tanzanian troops in Uganda for several years after Amin's overthrow, together with Nyerere's long-standing friendship with Uganda's former president, Milton Obote, also led to strained relations with some of Uganda's leaders as well as arousing suspicions in Kenya. Elsewhere in Africa, however, Nyerere was able to play an authoritative role, notably in the negotiations leading to the independence of Zimbabwe and in the formation of an organization of African states to try to resist economic domination by South Africa. (K.In.)

Events in Zanzibar caused continuing concern for the mainland leadership. The arbitrary arrest and punishment of anyone believed to oppose the state gave rise to regret that the constitution of the joint republic prevented the mainland authorities from intervening in the island's affairs where questions of law and justice were involved. The failure to hold elections in Zanzibar also contrasted unfavourably with developments on the mainland. In April 1972 Karume was assassinated by members of the military. His successor, Aboud Jumbe, had been a leading member of Karume's government, and, while his policies did not differ markedly from those of Karume, they appeared to be moving gradually closer into line with mainland practices. The amalgamation of TANU and the ASP under the title of Revolutionary Party (Chama cha Mapinduzi; CCM) early in 1977 was a hopeful sign but was followed by demands for greater autonomy for Zanzibar. This trend was checked when Ali Hassan Mwinyi succeeded Jumbe in 1984 and became president of the joint republic after Nyerere resigned in November 1985; however, in the late 1980s dissent resurfaced in Zanzibar, culminating in the revelation in January 1993 that Zanzibar had joined the Organization of the Islāmic Conference. Criticism on the mainland forced its withdrawal later that year.

The Revolutionary Party

Mwinyi inherited an economy suffering from the country's lack of resources, the fall in world prices for Tanzanian produce, the rise in petroleum prices, and inefficient management. An acute shortage of food added still further to his problems. Though he promised to follow Nyerere's policy of self-reliance, Mwinyi soon concluded that his predecessor's resistance to foreign aid could no longer be sustained. In accepting an offer of assistance from the International Monetary Fund (IMF) in 1986, Mwinyi adopted some structural reforms and furthered the devaluation of the currency begun in 1984 by Nyerere, who also had denationalized the state-run sector of the sisal industry in 1985. Moreover, private enterprise had been allowed to take over other areas of business. In May 1992 the constitution was amended to provide for a multiparty political system, and in 1995 the first national elections under this system were held. Nyerere's forecast that freedom would impose a heavy burden on the country had been realized, but his program of maximum self-reliance had failed to provide an immediate remedy for his country's problems. (K.In./Ed.)

For later developments in the history of Tanzania, see the BRITANNICA BOOK OF THE YEAR.

Uganda

The landlocked and mostly rural Republic of Uganda lies across the Equator approximately between latitudes 4° N and 1° S and longitudes 30° E and 35° E. Covering a total area of 93,070 square miles (241,000 square kilometres), the country is roughly the size of its former colonial ruler, Great Britain. Its neighbours are The Sudan to the north, Kenya to the east, Tanzania and Rwanda to the south, and Zaire to the west. The capital city, Kampala, is built around seven hills not far from the shores of Lake Victoria, which forms part of the frontier with Kenya and Tanzania.

Uganda obtained formal independence on Oct. 10, 1962. Its borders, drawn in an artificial and arbitrary manner in the late 19th century, encompassed two essentially different types of society: the relatively centralized Bantu kingdoms of the south and the Nilotic and Sudanic "tribes without rulers" to the north. This split between north and south—aggravated by uneven development under the British, which favoured the south—set the stage for years of turmoil in the postcolonial era, when, under military rulers such as Idi Amin, Uganda was torn by tyranny and bloodshed. When the violence subsided in the 1980s, the country faced the task of rebuilding, so that, with its beautiful green landscape and fertile agriculture, it could once again become what Winston Churchill called "the pearl of Africa." (O.H.K.)

Split between north and south

PHYSICAL AND HUMAN GEOGRAPHY

The land. *Relief.* Most of Uganda is situated on part of the Central Plateau, an enormous monotonous expanse that drops gently from about 5,000 feet (1,500 metres) in the south to approximately 3,000 feet in the north. The limits of Uganda's plateau region are marked by mountains and valleys.

To the west a natural boundary is composed of the Virunga (Mufumbiro) Mountains, the Ruwenzori Range, and the Western Rift Valley. The volcanic Virunga Mountains rise to 13,540 feet at Mount Muhavura and include Mount Sabinio (11,960 feet), where the borders of Uganda, Zaire, and Rwanda meet. Farther north the Ruwenzori Range—popularly believed to be Ptolemy's Mountains of the Moon—rises to 16,795 feet at Margherita Peak; its heights are often hidden by clouds, and its peaks are capped by snow and glaciers. Between the Virunga and Ruwenzori mountains lie Lakes Edward and George. The rest of the boundary is composed of the Western Rift Valley, which contains Lake Albert and the Albert Nile River.

The northeastern border of the plateau is defined by a string of volcanic mountains, including Mounts Zulia (7,048 feet), Morungole (9,022 feet), Moroto (10,116 feet), and Kadam (Debasien; 10,067 feet). The southernmost mountain—Mount Elgon—is also the highest of the chain, reaching 14,178 feet. South and west of these mountains is an eastern extension of the Rift Valley, as well as Lake Victoria. To the north the plateau is marked on the Sudanese border by the Imatong Mountains, with an elevation of about 6,000 feet.

Drainage. The country's drainage system is dominated by six major lakes—Victoria (26,828 square miles), the

world's second largest inland freshwater lake (after Lake Superior in North America), to the southeast; Edward and George to the southwest; Albert to the west; Kyoga in central Uganda; and Bisina in the east. Together with the lakes, there are eight major rivers. These are the Victoria Nile in central Uganda; the Achwa, Dopeth-Okok, and Pager in the north; the Albert Nile in the northwest; and the Kafu, Katonga, and Mpongo in the west.

The Nile system

The southern rivers empty into Lake Victoria, the waters of which escape through Owen Falls near Jinja and form the Victoria Nile. This river flows northward through the eastern extension of Lake Kyoga. It then turns west and north to drop over Karuma Falls and Murchison (or Kabalega) Falls before emptying into Lake Albert.

Lake Albert is drained to the north by the Albert Nile, which is known as the Baḥr al-Jabal, or Mountain Nile, after it enters The Sudan at Nimule. Rivers that rise to the north of Lake Victoria flow into Lake Kyoga, while those that rise north of Kyoga tend to flow into the Albert Nile. The rivers of the southwest flow into Lakes George and Edward.

Except for the Victoria and Albert Niles, the rivers are sluggish and often swampy. Clear streams are found only in the mountains and on the slopes of the Rift Valley. Most of the rivers are seasonal and flow only during the wet season, and even the few permanent rivers are subject to seasonal changes in their rates of flow.

Soils. The soils are predominantly ferralites (soils containing iron and aluminum). Interspersed with these are the waterlogged clays characteristic of the northwest and of the western shores of Lake Victoria. In general the soils are fertile, although they are of poorer quality in the north than in the south.

Climate. With the equator passing through southern Uganda, the climate is characteristically tropical. Temperatures are modified, however, by elevation and, locally, by the presence of the lakes. The major air currents are northeasterly and southwesterly. There is little variation in the sun's declination at midday, and the length of daylight is nearly always 12 hours. All of these factors, combined with an equatorial cloud cover, ensure an equable climate throughout the year.

Most parts of Uganda receive adequate rainfall; amounts range from a low of less than 20 inches (500 millimetres) a year in the northeast to a high of 80 inches in the Sese Islands of Lake Victoria. In the south there are two wet seasons, occurring in April and May and in October and November, which are separated by dry periods broken by tropical thunderstorms. In the north the climate is roughly divided into a wet season from the months of April to October and a dry season that lasts from November to March.

Plant and animal life. Wooded savanna (grassy parkland) is typical of central and northern Uganda. Under less favourable conditions, dry acacia woodland, dotted with the occasional candelabra (tropical African shrubs or trees with huge spreading heads of foliage) and euphorbia (plants often resembling cacti and containing a milky juice) and interspersed with grassland, occurs in the south. Similar components are found in the vegetation of the Rift Valley floors. The steppes (treeless plains) and thickets of the northeast represent the driest regions of Uganda. In the Lake Victoria region and the western highlands, the mosaic of elephant grass and forest remnants appears to have resulted from human incursions affecting the former forest covering. The medium-elevation forests contain a rich variety of species, with many representatives of West African vegetation. The high-elevation forests of Mount Elgon and the Ruwenzori Range occur above 6,000 feet; on their upper margins they give way, through transitional zones of mixed bamboo and tree heath, to high mountain moorland. Uganda's 5,600 square miles of swamp include both papyrus swamp and seasonal, grassy swamp.

Remnants of forest

Lions and leopards are widely distributed but are seen only infrequently. Hippopotamuses and crocodiles inhabit most lakes and rivers, although the latter are not found in Lakes Edward and George. Mountain gorillas, chimpanzees, and small forest elephants occur only in the extreme west. Elephants, buffalo, and the Uganda kob (an antelope) are found in the west and north, while the black rhinoceros, white rhinoceros, and giraffe are confined to the north. Zebras, topis, elands, and roan antelopes occur both in the northeastern and southern grasslands, but other kinds of antelopes, such as the oryx, greater and lesser kudu, and Grant's gazelle, live only in the northeastern area. The varied fish life includes ngege (a freshwater, nest-building fish of the tilapia species) and Nile perch.

Insects are a significant element in the biological environment. The female anopheles mosquito may transmit malaria anywhere below 5,000 feet, and extensive areas of good grazing are closed to cattle because of the presence of deadly tsetse flies.

The three national parks contain an interesting variety of animal life. Murchison Falls (or Kabalega) National Park, with an area of 1,500 square miles, stretches on either side of the Victoria Nile. Queen Elizabeth (or Ruwenzori) National Park occupies some 850 square miles in the Lake Edward–Lake George basin, and Kidepo National Park consists of 480 square miles of magnificent country adjacent to the Sudanese frontier. (M.S.Ki.)

Settlement patterns. Uganda's population remains basically rural, although the urban population is growing at a faster rate. Virtually all of Uganda's cities and towns arose from the colonial experience. The capital, Kampala, is the largest, and the other major cities are Jinja, Mbale, Mbarara, Masaka, Entebbe, and Gulu. Apart from Gulu, all of these are located in the south, where centralized and hierarchical political communities, often led by a king, existed before colonization.

The growth of urban centres has been caused by a rural–urban migration within the south itself as well as from the north to southern towns. These towns resemble Western cities more than they do rural Uganda. Before independence, urban dwellers were mainly British and Asian immigrants; only gradually did a minority of black urbanites begin to emerge, and then Idi Amin's "economic war" of the 1970s expelled most immigrants, who were replaced by black Ugandans. Since then, an interesting consequence has been the "ruralization" or "villagization" of towns, as formerly open spaces and parks have been converted into private little gardens. To some extent, in some instances town and country have merged.

Surrounding most urban centres are "shantytowns," dirty, ugly, and crowded neighbourhoods inhabited by the very poor who have abandoned rural life and yet do not fit into the cities. These slums, whose inhabitants include informal petty traders, prostitutes, pimps, manual labourers, and pickpockets, are often breeding grounds for crime, violence, and disease.

The rural majority is as diverse as the number of Ugandan ethnic groups. However, they can be divided into two groups: the nomadic pastoral Ugandans and the relatively more sedentary farmers. These rural peoples are the custodians of Uganda's indigenous cultures, yet since 1900 modernization has been creeping in on them, too, as roads, electricity, piped water, and health-care centres have gradually reached the countryside. Historically, rural life has been kinship-based, with people living in intimate communities of broadly defined relatives, but the relative modernization and mechanization of agriculture, including the introduction of modern ranching and cash crops, have helped to transform rural life.

"Nilotic North" and "Bantu South"

The people. *Ethnic groups.* Although Uganda is inhabited by a large variety of ethnic groups, a division is usually made between the "Nilotic North" and the "Bantu South." Bantu speakers represent over 70 percent of Uganda's population. Of these, the Ganda remain the largest single ethnic group, constituting about 20 percent of the total national population. In addition to the Ganda, other members of the so-called Eastern Lacustrine wing of the Bantu are the Soga, Gwe, Gisu, Nyole, Samia, and Kenyi. The Western Lacustrine branch includes the Nkole, Toro, Nyoro, Kiga, Amba, and Konjo.

Like the Bantu South, the Nilotic North is often divided into western and eastern branches. The Western Nilotes represent approximately 15 percent of Uganda's population and are usually known collectively as the Luo. The Acholi and Lango are the two largest subgroups of this

category; smaller groups are the Alur, Padhola, Kumam, and Jonam. The Eastern Nilotes constitute approximately 12 percent of Uganda's population. Their four largest members are the Dodoth, Teso, Jie, and Karamojong; closely related is a variety of smaller groups including the Kakwa, Sebei, Labwor, Nyakwai, Tepeth, Napore, Nyangea, and Teuso. Also to be found in the north are Central Sudanic peoples, who include the Lendu, Lugbara, and Madi. Together they constitute about 6 percent of Uganda's population.

Under British colonial rule, economic power and education were concentrated in the south. As a result, the Bantu came to dominate modern Uganda, occupying most of the high academic, judicial, bureaucratic, and religious positions and a whole range of other prestigious roles. On the other hand, the armed forces created by the British recruited overwhelmingly from the north, and the police, paramilitary forces, and prisons were also dominated by northerners. This meant that, just as economic power lay in the south, military power was concentrated in the north. The political events of postcolonial Uganda have to a large extent been shaped by this imbalance.

By 1969 there were about 75,000 Indians, Pakistanis, and Bangladeshis in Uganda, most of whom lived in cities and towns and were primarily involved in commerce and trade. Although Ugandan citizenship was available to them, most Asians preferred to retain British passports. Since the sudden expulsion in 1972 by Idi Amin of all noncitizen Asians, the Asian population has almost disappeared from Uganda. Amin's action proved immensely popular with most indigenous Ugandans, but the country has yet to recover completely from the economic consequences of those swift expulsions.

In the 1960s Uganda also had approximately 10,000 resident western Europeans and North Americans, who served mostly in the professions. Most left the country around the time of the Asian deportations.

Languages. The language of official business in Uganda is English. Although only a tiny fraction of the populace speaks English well, access to high office, prestige, and economic and political power is almost impossible without an adequate command of this language. Swahili has been chosen as another official national language, and its potential for facilitating regional integration is immense. However, the command of Swahili by most Ugandans falls substantially below that of peoples in Tanzania, Kenya, and even eastern Zaire.

The major languages

Uganda's indigenous languages are coextensive with its different ethnic groups. Only the major native languages possess a sustained or extensive written tradition. In addition to English, French, Arabic, and Swahili, Radio Uganda broadcasts in the following indigenous languages: Alur, Ateso, Dhopadhola, Kakwa, Karamojong, Kumam, Kupsabiny, Luganda, Lugbara, Lugwe, Lumasaba, Lunyole, Luo, Lusamia, Lusoga, Madi, Rukiga, Rukonjo, Runyankole, Runyoro, and Rutoro. Most Ugandans can understand at least one of these.

Religions. Uganda has a triple religious heritage of indigenous religions, Islām, and Christianity. Reliable statistics are rare, but most figures suggest that the largest religious community is that of the Roman Catholics (about one-half of the population), followed by that of the Anglicans with one-quarter of the population and the Muslims with less than 10 percent.

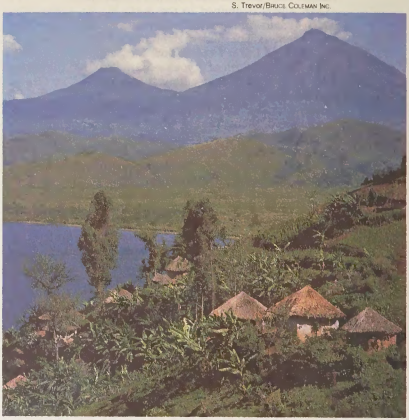
Of the immigrant religions, Islām was the first to arrive, but its believers remain a small minority, and it was politically significant only under Amin. In the colonial period Uganda witnessed a spirited Christian missionary activity—especially in the south, where Catholics were called *bafaransa* ("the French") and Protestants *bangerezza* ("the British"). Rivalry and even hostility between adherents of these two branches of Christianity, which have always been sharper and deeper than those between Christians and Muslims, are still alive today.

Although the immigrant religions dominate official statistics, most knowledgeable observers agree that the numerous indigenous religions command the greatest number of adherents. This is because most Ugandans, even after converting to Islām or Christianity, do not abandon their traditional beliefs altogether but blend them with their adopted faiths.

Demographic trends. With an annual growth rate of about 3 percent, the doubling time for Uganda's population is only a little more than 20 years. As many as one-half of Ugandans are under 15 years of age.

Barely 10 percent of Ugandans reside in cities or towns. Of these centres the most important are Kampala, the political and commercial capital, where approximately 35 percent of Uganda's urban population is concentrated, and Jinja, the industrial capital. The most densely populated areas are in the south, especially around Lake Victoria and Mount Elgon.

S. Trevor/Bruce Coleman Inc.



Houses near Mount Muhavura, western Uganda.

The economy. The economy is basically agricultural, with Uganda's moderate climate congenial to the production of both livestock and crops.

As has been the case with most African countries, economic development and modernization have remained elusive—a predicament that has been exacerbated by the country's chronic civil wars and sometimes by drought. In order to repair the damage done to the economy under Idi Amin, the government has encouraged foreign investments in agriculture and core industries, mainly from Western countries and former Asian residents. Attempts have also been made to secure support and loans from the World Bank and the International Monetary Fund.

Agriculture. Agriculture accounts for more than 95 percent of Uganda's export earnings and half of the gross domestic product (GDP), and it is the main source of income for well over 90 percent of the adult population. More than half of agricultural production is done by subsistence farmers, the vast majority of whom are based in the south, where there is more rainfall and fertile soil. Small-scale mixed farming predominates. The two most important cash crops for export are coffee and cotton. Tea and sugar are also grown for export. Food crops include corn (maize), millet, sorghum, cassava, sweet potatoes, plantains, peanuts (groundnuts), and various vegetables.

Major crops

Industry. Industry contributes less than 10 percent of the GDP and employs a commensurate portion of the labour force. The major industries are in the processing of such agricultural products as tea, tobacco, sugar, coffee, cotton, animal feeds, grains, dairy products, and edible oils, in the brewing of beer, and in the manufacture of cement, fertilizers, matches, metal products, paints, shoes, soap, steel, textiles, and motor vehicles.

Because of a serious lack of technical and management

personnel, a scarcity of spare parts, machinery, and fuel, and not least because of the devastation of war, industrial production has declined immensely since the early 1970s. By contrast, from 1961 to 1970 industrial output grew by 7.8 percent annually. With the return of stability to the country, foreign companies and lending institutions have invested in textile and steel mills, a motor vehicle assembly plant, a tannery, bottling and brewing plants, and cement factories.

Finance. Uganda's central bank, the Bank of Uganda, was founded in 1966. It monitors Uganda's commercial banks, serves as the government's bank, and issues the national currency, the Uganda shilling. The government sets the shilling's official exchange rate against foreign currencies, but in the "parallel economy," or black market, it can be obtained for a fraction of its official value.

The Uganda Commercial Bank, the Uganda Cooperative Bank, and the Uganda Development Bank serve most of the commercial and financial needs of the country. There are also commercial banks owned by British, Indian, and Libyan firms.

Trade. Uganda's principal exports are coffee, tea, cotton, copper, and sugar. Skins and hides are also sold outside the country. Coffee alone accounts for more than 90 percent of Uganda's export earnings. The principal destinations are the United States, France, Germany, Spain, the United Kingdom, and Japan.

Uganda imports machinery and transport equipment, assorted manufactured goods (including various metal products), fuels, and foodstuffs. The principal sources of imports include Kenya and Tanzania, Germany, the United Kingdom, and India.

Transportation. Being a landlocked state, Uganda relies heavily on Kenya and Tanzania (particularly the former) for access to the sea. Linking Kampala with Kilindini Harbour at Mombasa in coastal Kenya is a rail line that passes via Jinja, Tororo, Leseru, Nakuru, and Naivasha. Kampala is also connected to the north by a rail line that traverses the Pakwach bridge and to the western parts of the country by a line that reaches the border town of Kasese.

The state-owned Uganda Airlines offers international as well as domestic flights. The main international airport is at Entebbe, Uganda's former capital, about 20 miles to the west of Kampala.

Less than one-fourth of Uganda's road system is paved. There are river services on the Kagera River and lake services on Lakes Albert and Victoria.

Administration and social conditions. *Government.* Until 1967 Uganda was a quasi-federal polity that included five subregional monarchies, nonmonarchical districts, and a central government. The republican constitution adopted in 1967 abolished the monarchies and assigned ultimate political power to an elected president. The president was to be aided by a ministerial cabinet drawn, in the British tradition, from among members of the unicameral National Assembly. In theory the judiciary, legislature, and executive were to be autonomous, if coordinate, institutions of governance, but in reality the powers of the different branches of government varied widely with each president. Under Idi Amin's presidency (1971–79), representative institutions were abolished altogether, and, with the first of several military coups d'état in 1985, the constitution was suspended. While a commission drafts a new constitution, power rests in the hands of the president. A legislature called the National Resistance Council is composed of members appointed by the president as well as members elected by district and county Resistance Committees. Political parties exist, but active campaigning during elections is forbidden.

The 1967 constitution

Uganda is divided into 10 provinces: Busoga, Eastern, Kampala, Karamoja, Nile, North Buganda, Northern, South Buganda, Southern, and Western. The president appoints the provincial governors, who in turn appoint commissioners to run the 33 districts.

Education. Many of the oldest schools in Uganda were established by Christian missionaries from Europe. Since independence their role has been superseded by that of the government, but, because the number of secondary schools is modest, mission schools remain an important component of Uganda's educational system.

Makerere University in Kampala, which began as a secondary school in the 1920s, was the first major institution of higher learning in East and Central Africa. In addition to its medical school, Makerere's faculties include those of agriculture, arts, education, engineering, law, science, social sciences, and veterinary science. Other institutions previously affiliated to Makerere are the Institute of Teachers' Education and the National College of Business Studies.

Health. The vast majority of Uganda's hospitals are government-operated, the remainder being run by private charitable organizations.

The major killer diseases are malaria, measles, sleeping sickness, venereal diseases, shigellosis (dysenterial infections), whooping cough, hookworm, poliomyelitis, schistosomiasis, yaws, smallpox, tuberculosis, chicken pox, typhoid, and leprosy. Unclean water is a major carrier of these diseases. AIDS (acquired immune deficiency syndrome), known locally as "slim" because of its debilitating effects, is spreading at an alarming rate, with fully 10 percent of all adults said to be infected with the AIDS virus.

In the late 1980s the infant mortality rate per 1,000 live births was 103. Average life expectancy was 49 years for men and 53 years for women. Besides disease, the death rate in Uganda has been affected by years of anarchy, tyranny, economic dislocation, and the disintegration of government medical services.

Cultural life. *Daily life.* In spite of Uganda's ethnic diversity, virtually all communities are both patriarchal (male-dominated politically) and patrilineal (with descent traced through the father). Polygyny is widespread, while polyandry is unheard of. Being basically a rural people, most Ugandans, male and female, spend their daily lives working the land. Most children now go to school, but where a family choice must be made about which child to educate, it is often the son and rarely the daughter who is favoured.

The staple diet in most of the south is a kind of plantain called *matoke.* Often sweet potatoes and cassava, too, are consumed. The bulk of the north consumes millet, sorghum, cornmeal, and cassava. The pastoralist communities tend to consume animal-derived products, especially butter, meat, and animal blood. Fish is eaten throughout the country by all groups.

Ugandan foods

The most popular sport by far is association football (soccer). Other popular sports and games include boxing, wrestling, basketball, field hockey, rugby, athletics (track and field), netball, and golf.

The arts. The Westernized elites are virtually the sole consumers and practitioners of the fine arts. However, indigenous music remains popular and is widely broadcast by Radio Uganda and Uganda Television. The most popular foreign tunes are Zairean and Western music. Most nightclubs and dance halls often combine all of these musical traditions.

In precolonial times there were no individual works of art, since creativity tended to be basically collective and cumulative. Under European influence, however, Uganda now has developed well-known individual musicians, dancers, writers, actors, and other artists.

Uganda has nine museums, the largest and most important being the Uganda Museum in Kampala. Other museums include those at Murchison Falls and Queen Elizabeth national parks. The National Theatre is based in Kampala, while a wide variety of smaller theatres exist across the country.

The news media. Radio Uganda is government-owned, as is Uganda Television. Television is transmitted over a radius of 200 miles from Kampala, and there are relay stations around the country.

A fluctuating range of daily newspapers is published in Uganda, all of them in Kampala. Daily newspapers published in English include *Telecast, The Star,* and the state-owned *New Vision.* There are also three dailies in Luganda (*Taifa Empya, Munno,* and *Ngabo*) and various weeklies, periodicals, and assorted papers in both English and Luganda.

The degree of governmental control and censorship of the press has varied under different regimes, but, as elsewhere in the Third World, a completely free and independent press has never existed in Uganda. (O.H.K.)

For statistical data on the land and people of Uganda, see the *Britannica World Data* section in the BRITANNICA BOOK OF THE YEAR.

HISTORY

There are no written records for the history of Uganda dating back much before the mid-19th century. For some centuries before that, however, Nilotic and Nilo-Hamitic tribes had been migrating southward into the area of Lake Victoria, where they had blended with or superimposed themselves upon the Bantu inhabitants. Possibly in the 15th century a race of tall, fair-skinned, and finely featured Hima people, cattle keepers, moved into what is now western Uganda and became the overlords of the area, treating as servants the agricultural peasantry who inhabited the country before them. This invasion was followed by another wave, this time of Nilotes, from around the Al-Ghazāl River in the Sudan. These latter, as they crossed the Nile, pushed the earlier invaders before them. Their advance guard settled in western Uganda and established the state of Bunyoro-Kitara. Other offshoots of the same movement settled in Acholi, while some recrossed the Nile northward to make their homes for a time in what is now Lango district. The advance of the Lango themselves forced the earlier migrants to resume their journey eastward and southward until they finally took up their residence on the northeastern shore of Lake Victoria in the present Nyanza region of Kenya. This move took place only in the 19th century, for the Lango, Teso, and other Nilo-Hamitic groups who began to move into what is now northern Uganda toward the end of the 17th century seem to have halted in their advance for almost 100 years and recommenced their southward march to the Nile only toward the end of the 18th century.

Nilotic invasion

Bunyoro and Buganda. The organization of the tribes who came to inhabit the area north of the Nile was mainly based upon their clan structure. In this respect the northern tribes differed markedly from the peoples to the southwest of the Nile, where the dominant state was that of Bunyoro-Kitara, which, under able rulers, extended its influence eastward and southward over a considerable area. To the south there were a number of lesser Hima states, each with its chief, who, like the ruler of Bunyoro-Kitara, combined priestly functions with those of a secular leader. To the southeast of Bunyoro-Kitara the smaller state of Buganda grew up as an offshoot of its larger neighbour. By the end of the 18th century, however, the boundaries of Bunyoro-Kitara had been stretched so far that the authority of the ruler began to crumble, and a succession of pacific chiefs accelerated this decline. Simultaneously the smaller, more compact state of Buganda enjoyed a succession of able and aggressive kabakas, or rulers, and began to expand at the expense of Bunyoro-Kitara.

It was during this period of Buganda's rise that the first Arab traders reached the country in the 1840s. Their object was to trade in ivory and slaves. Mutesa I, who took office about 1856, admitted the first European explorer, Captain John Hanning Speke, who crossed into the kabaka's territory in 1862. Speke was not impressed by Mutesa, whom he regarded as an arrogant youth.

Henry Morton Stanley, who reached Buganda in 1875, found Mutesa a more mature man. Experience and the dangers facing his country had mellowed his character. Although Buganda had not been attacked, Acholiland, to the north, had been ravaged by slavers from Egypt and the Sudan since the early 1860s, and, on the death of Kamrasi, the ruler of Bunyoro, his successor, Kabarega, had defeated his rivals only with the aid of the slavers' guns. Moreover, an emissary from the Egyptian government, Linant de Bellefonds, had reached Mutesa's headquarters just before Stanley, so that the kabaka was anxious to obtain allies. He readily agreed to Stanley's proposal to ask Christian missionaries to go to Uganda but was disappointed, when the first agents of the Church Missionary Society arrived in 1877, to find that they had no interest in military matters.

Christian missions

In 1879, representatives of the Roman Catholic White Fathers Mission also reached Buganda, and Mutesa was shrewd and strong enough to ensure that the missionaries did not stray far from his headquarters. Nonetheless, their influence rapidly increased because they were in constant touch with the chiefs whom the kabaka kept around him, and inevitably they became drawn into the politics of the country. Mutesa himself had little to fear from these new influences, and when, owing to the Mahdist rising in the Sudan, Egyptian expansion was checked, he felt confident in dealing brusquely with the handful of missionaries in his country. His successor, Mwanga, who became kabaka in 1884, was a weaker character and soon became jealous of their influence. Attempting to drive the missionaries and their supporters from the country, he was himself deposed in 1888, only to be followed soon after by his intended victims, who had been forced to flee by Muslim Ganda, whose numbers had grown under Arab influence.

The Uganda Protectorate. Mwanga's restoration with the assistance of the Christian Ganda, both Roman Catholic and Protestant, coincided with the first contact of European imperialism with Buganda. The German adventurer Carl Peters made a treaty of protection with Mwanga in 1889, but this was revoked when the Anglo-German agreement of 1890 declared all the country north of latitude 1° S to be a British sphere of influence. The Imperial British East Africa Company agreed to administer the region on behalf of the British government, and in 1890 the company's agent, Captain F.D. Lugard, made a treaty with Mwanga, placing Buganda under the company's protection. Lugard also made treaties of protection with two other chiefs, the rulers of the western states of Ankole and Toro. Shortage of money soon forced the company to abandon its charge, however, and, partly for strategic reasons and partly as a result of pressure from missionary sympathizers in Britain, the British government itself declared a protectorate over Buganda in 1894.

The government inherited from the company a country divided into politico-religious factions that, in Lugard's time, had even erupted into civil war (1892). Buganda was also threatened by Kabarega, the ruler of Bunyoro, but an expedition against him in 1894 deprived him of his headquarters and made him a refugee for the rest of his career in Uganda. The protectorate was extended in 1896 to include Bunyoro, Toro, Ankole, and Busoga, and in the later 1890s treaties were also made with chiefs to the north of the Nile. Mwanga revolted against British overlordship in 1897 and was overthrown and replaced by his infant son.

A mutiny of the administration's Sudanese troops in 1897 roused the British government to take a more active interest in the Uganda protectorate, and in 1899 Sir Harry Johnston was commissioned to visit the country and to make recommendations on its future administration. The main outcome of his mission was the Buganda Agreement of 1900, which formed the basis of British relations with Buganda for more than 50 years. Under its terms the kabaka was recognized as ruler of Buganda as long as he remained faithful to the protecting authority. His council of chiefs, the *lukiko,* was given statutory recognition. It was the leading chiefs who benefited most from the agreement, since in addition to the greater authority that they acquired, they were also granted land in freehold to ensure their support for the negotiations. Johnston made another agreement of a less detailed nature with the ruler of Toro (1900), and subsequently a third agreement was made with the ruler of Ankole (1901). Meanwhile British administration was being gradually extended north and east of the Nile, but in these areas, where a centralized authority was unknown, no agreements were made and British officers, frequently assisted by Buganda agents, administered the country directly. By 1914 Uganda's boundaries had been fixed and administration had reached most areas.

The Buganda Agreement of 1900

Growth of a peasant economy. Early in the 20th century Sir James Hayes Sadler, who succeeded Johnston as commissioner, concluded that the country was unlikely to prove attractive to European settlers. Sadler's own successor, Sir Hesketh Bell, turned this practical observation into a dogma and announced that he wished to develop Uganda

as an African state. In this he was opposed by a number of his more senior officials and in particular by the chief justice, William Morris Carter. Carter was chairman of a land commission whose activities continued until after World War I. Again and again the commission urged that provision be made for European planters, but their efforts were unsuccessful. Bell himself had laid the foundations of a peasant economy by encouraging the cultivation of cotton, which had been introduced into the country as an economic crop in 1904. It was mainly owing to the wealth derived from cotton that Uganda became independent of a grant-in-aid from the British Treasury in 1914.

Although there were a few skirmishes on the southwestern frontier in 1914, Uganda was never in danger of invasion, but World War I did retard the country's development. Soon after the war it was decided that the protectorate authorities should concentrate, as Bell had suggested, upon encouraging African agriculture. This policy was made necessary by the British government's decision to forbid the alienation of land in freehold; and the depression of the early 1920s dealt a further blow to the hopes of European planters. Henceforward, plantation-grown coffee, though second in importance to cotton for the time being, never threatened the supremacy of Uganda's chief product. The part to be played by Europeans as well as Asians was now mainly on the commercial and processing side of the protectorate's agricultural industry. When an additional crop was sought in order to widen the basis of the economy, it was African-grown coffee that the agricultural department encouraged.

The increased output of primary produce called for a considerable extension and improvement in communications. Just before World War I a railway had been built running northward from Jinja, on Lake Victoria, to Namasagali, with a view to opening up the Eastern Province.

Rail link to the Kenyan coast

In the 1920s the railway from Mombasa, on the Kenyan coast, was extended to Soroti, and in 1931 a rail link was also completed between Kampala, the industrial capital of Uganda, and the coast.

The depression of the early 1930s caused a halt in Uganda's economic progress, but the protectorate's recovery was more rapid than that of its neighbours, so that the later 1930s were a period of steady expansion.

Political and administrative development. In 1921 a Legislative Council was instituted, but its membership was so small (four official and two nonofficial members) that it made little impact upon the country. The Indian community, which played an important part in the commercial life of the protectorate, was resentful to find that it was not to have equal representation with Europeans on the unofficial side of the council and so refused to participate until 1926. There was no evidence of a desire on the part of the Africans to sit in the council, since the most politically advanced section of the community, the Ganda, regarded its own *lukiko* as the most important council in the country.

In view of this attitude on the part of the Africans toward the protectorate legislature, it is not surprising that they opposed the suggestion, current in the later 1920s, that there should be some form of closer union between the East African territories. It was not only an interest in tribal traditions but also a fear of domination by Kenya's European settlers that stimulated this opposition. The Europeans of the protectorate also resented what they regarded as the subordinate character of Uganda's relations with Kenya, while the Asians were as worried as the Africans about the possible supremacy of Kenya's white settlers.

Perhaps the main feature in local government between the wars was the gradual replacement of older chiefs, men of strong personality but usually lacking a formal education, by younger, better educated men, more capable of carrying out the protectorate administration's policy and more amenable to British control. In Buganda, too, the protectorate administration began to interfere more actively in the kingdom's affairs in order to increase efficiency. The main result was to lessen the respect shown by the people to the chiefs outside Buganda and to cause resentment on the part of some of the Ganda chiefs against the insidious curtailment of their powers. Another important development was the beginning of government interest in education. Missionary societies had opened a number of good schools in Buganda. In 1925 an education department was set up by the protectorate administration, and, while aid was given to the missionary societies, government schools were also opened.

World War II and its aftermath. In World War II Uganda was faced with the task of becoming as self-sufficient as it could. More important than the actual military struggle as far as Uganda was concerned, however, was an attempt by the governor, Sir Charles Dundas, to reverse his predecessors' policy and to give more freedom to the factions striving for power in Buganda. After an outbreak of rioting in 1945, however, the old policy was revived. In 1945, too, the first Africans were nominated to the Legislative Council, and in succeeding years there was a steady increase in African representation. An important step was taken in 1954 when the African membership of the council was increased to 14 out of a total of 28 nonofficial members; the 14 were selected to represent districts thought to be more natural units of representation than the provinces that had previously been represented. In 1955 a ministerial system was introduced, with 5 nonofficial ministers out of a total of 11. The success of the council was undermined, however, by Buganda's erratic participation, since a central legislature was thought to threaten the degree of autonomy that Buganda enjoyed. This feeling was strengthened by the deportation in 1953 of Mutesa II, the kabaka, for refusing to cooperate with the protectorate government. He returned in 1955 as a constitutional ruler, but the rapprochement between Buganda and the protectorate government was lukewarm.

Africans on the Legislative Council

In the immediate postwar years the protectorate administration placed greater emphasis upon economic and social development than upon political advance. From 1952 there was a rapid increase in government provision for secondary education, while legislation was enacted and a loan fund established to encourage Africans to participate in trade. A relatively ambitious development program was greatly assisted by the high prices realized for cotton and for coffee, which overtook cotton as Uganda's most valuable export in 1957. In 1954 a large hydroelectric project was inaugurated at Owen Falls on the Nile near Jinja, and in 1962 a five-year development plan was announced.

The Republic of Uganda. In the late 1950s, with the emergence of a few political parties, the attention of the African population was concentrated on achieving self-government, though, owing to the poor organization of the parties and the fact that they were based upon Buganda, the focus of interest remained in the Legislative Council, where the other parts of the country felt themselves to be better represented. The kingdom of Buganda intermittently pressed for independence from Uganda, and the question of the protectorate's future status became acute. Discussions in London in 1961 led to full internal self-government in March 1962. Benedicto Kiwanuka, a Roman Catholic Ganda who was formerly chief minister, became the first prime minister, but in the elections in April 1962 he was displaced by Milton Obote, a Lango who headed the Uganda People's Congress party (UPC). At further discussions in London in June 1962, it was agreed that Buganda should receive a wide degree of autonomy within a federal relationship. Faced with the emergence of Obote's UPC, which claimed support throughout the country apart from Buganda, and of the Democratic Party (DP), which was based in Buganda and led by Kiwanuka, conservative Ganda leaders set up their own rival organization, Kabaka Yekka (KY), "King Alone."

Obote's first presidency. Divided politically on a geographic as well as an ethnic basis, Uganda became independent on Oct. 9, 1962. By accepting a constitution that conceded what amounted to federal status to Buganda, Obote contrived an unlikely alliance with the Ganda establishment. Together the UPC and KY were able to form a government with Obote as prime minister and with the DP in opposition. In an attempt to weld the alliance more firmly, Obote agreed to the appointment of Mutesa II as the country's first president, replacing the British governor-general. The move was unsuccessful. Although

Conflict between prime minister and kabaka

Obote was able to win over some of the members of the KY and even of the DP so that they joined the UPC, tension between the kabaka on the one hand and the UPC on the other grew steadily. The Ganda leaders particularly resented their inability to dominate a government composed mainly of members of other ethnic groups. Inside the UPC there were also divisions, because each member of parliament owed his election to local ethnic supporters rather than to his membership in a political party, and those supporters frequently put pressure upon their representatives to redress what they saw as an imbalance in the distribution of the material benefits of independence.

Faced with this dissatisfaction among some of his followers and with increasingly overt hostility in Buganda, Obote arrested five of his ministers and suspended the independence constitution in 1966. Outraged, the Ganda leaders ordered him to remove his government from the soil of the kingdom. Obote responded by sending troops under the leadership of Colonel Idi Amin to arrest the kabaka. The latter escaped to England, where he died in 1969, but the violence that accompanied the army's intervention fueled the anger of the Ganda still further. When Obote imposed a new republican constitution—appointing himself executive president, abolishing all the kingdoms, and dividing Buganda into administrative districts—he also lost the support of the peoples of southwestern Uganda. From this time internal friction grew in intensity, fostered by mutual suspicion between the rival groups, by assassination attempts against the president, and by the increasingly oppressive methods employed by the government to silence its critics.

At independence the export economy was flourishing without adversely affecting agricultural production for subsistence purposes, and the economy continued to improve, largely because of the demand and high prices paid for coffee. To meet accusations that the profits from exports did not benefit the producers to the extent that they should, Obote decided in 1969 to try to distribute the benefits arising from the prospering economy more widely. To this end he published a "common man's charter," aimed at removing the last vestiges of feudalism. The government, he said, would take a majority holding in the shares of the larger companies, mainly foreign-owned. In order to unite the country more firmly, he also produced a plan for a new electoral system in 1970 that would require successful candidates for parliament to secure votes in constituencies outside their home districts.

These proposals met with a cynical response in some quarters, and before they could be put into effect the government was overthrown. Obote had relied heavily upon the loyalty of Idi Amin, but the latter had been building support for himself within the army by recruiting from his own Kakwa ethnic group in the northwest. The army, theretofore composed of Acholi and their neighbours, Obote's own Lango people, now became sharply divided. Simultaneously, a rift developed between Obote and Amin, and in January 1971 the latter took advantage of the president's absence from the country to seize power.

Rise of Idi Amin

Tyranny under Amin. The coup was widely welcomed, notably by the Ganda, whose hatred for Obote made them well disposed toward any usurper. Several Western powers, including Britain, who feared the spread of communism, were also relieved at Obote's overthrow because they had become suspicious of what they saw as a tendency for his policies to move to the left.

Amin's inability to govern was soon revealed. Having had little education and virtually no officer training, he relied upon innate shrewdness and arbitrary violence to maintain himself in power. He destroyed the one potential centre of effective opposition by a wholesale slaughter of senior army officers loyal to Obote. To win more general support, he ordered all Asians who had not taken Ugandan nationality to leave the country in 1972. His move won considerable approval because many Africans believed that they had been exploited by the Asians, who controlled the middle and some of the higher levels of the country's economy. More perceptive critics recognized the difficulty of replacing Asian commercial expertise from internal resources, and their forebodings were quickly realized. Although a few wealthy Ugandans profited from their acquiescence and participation in Amin's actions, the majority of the commercial concerns formerly owned by Asians were given to senior army officers who rapidly squandered the proceeds and then allowed the businesses to collapse.

Most people in the countryside were able to ride out the total breakdown of the economy that took place in the middle and late 1970s because of the fertility of Uganda's soil. In the towns, an all-pervading black market developed, and dishonesty became the only means of survival. This economic and moral collapse stirred up criticism of the government, but any opposition was ruthlessly crushed by the army, and arbitrary violence, directed particularly against the better educated and therefore more critical members of the community, made the people afraid to rebel. This feeling of insecurity spread even to the leaders of Amin's army and weakened the president's hold on the country.

In an attempt to divert attention from Uganda's internal problems, Amin launched an attack on Tanzania in October 1978. It was the prelude to his downfall. Tanzanian troops, assisted by armed Ugandan exiles, quickly put Amin's demoralized army to flight. A coalition government of former exiles, calling itself the Uganda National Liberation Front (UNLF), with a former leading figure in the DP, Yusufu Lule, as president, took office on April 13, 1979. Disagreement over economic strategy and the fear that Lule was promoting the interests of his own Ganda people led to his being replaced in June by Godfrey Binaisa, but Binaisa's term of office was also short-lived. Supporters of Obote plotted his overthrow, and Obote returned to Uganda on May 27, 1980.

Instability after Amin

Obote's second presidency. In December 1980 Obote's party, the UPC, won a majority in highly controversial elections for parliament. The DP leadership reluctantly agreed to act as a constitutional opposition, but one man who had played a significant part in the military overthrow of Amin, Yoweri Museveni, refused to accept the UPC victory. Forming a guerrilla group in the bush near Kampala, he waged an increasingly effective campaign against the government.

With the support of the International Monetary Fund and other external donors, Obote tried hard to re-create the economy. Initially his efforts seemed to meet with some success, but two factors militated against him. First, the appalling inflation resulting from Amin's profligacy, and the black market to which it gave birth, made it impossible for urban wage earners to keep pace with rising prices. Salaried civil servants grew frustrated at the government's inability to increase their pay in line with their needs. Second, the guerrilla war drew strength from being based in Buganda, among people already suspicious of Obote. That strength grew as an ill-paid, ill-disciplined, and vengeful army, consisting largely of Acholi and Lango, ravaged the countryside for loot and took vengeance upon their longtime Ganda enemies.

Museveni in office. It was a split within the army itself—in particular, between its Acholi and Lango members—that led to Obote's overthrow and exile and to the seizure of power by an Acholi general, Tito Okello, but this could not prevent the victory of Museveni's force of southern fighters, now calling themselves the National Resistance Army (NRA). Museveni became president on Jan. 29, 1986, and the long period of rule by northerners came to an end. Pending the drafting of a new constitution, an indirectly elected National Resistance Council, dominated by the National Resistance Movement, acted as the national legislature.

Faced with the same problems that had confronted the UNLF in 1979 and Obote in 1980, problems exacerbated by the murderous struggle that had devastated one of the country's richest agricultural regions, Museveni announced a policy of moral as well as economic reconstruction. It was not easy to enforce. Sporadic military resistance to the new government continued, particularly in the north and east. Arms were plentiful, and dissatisfied persons were willing to use them to promote their ends. The NRA, despite the president's injunctions, sometimes proved as

Reconstruction

heavy-handed in dealing with opponents as Obote's forces had been. (K.In.)

Security did improve, however, and observers claimed that human rights were more widely protected. A constitution promulgated in 1995 led to the restoration of the local monarchies, and presidential elections were held in May 1996; Museveni easily won the majority of votes. In spite of continued economic growth, inflation and unemployment persisted, as did Uganda's dependence on fluctuating markets for its agricultural produce, and institutional corruption continued to be an issue in the 21st century. (Ed.)

For later develoments in the history of Uganda, see the BRITANNICA BOOK OF THE YEAR.

THE COUNTRIES OF THE HORN OF AFRICA

Djibouti

The Republic of Djibouti (French: République de Djibouti; Arabic: Jumhūrīyah Jībūtī), a strategically located nation on the northeast coast of the Horn of Africa, is situated on the Strait of Mandeb, which lies to the east and separates the Red Sea from the Gulf of Aden. Small in size (8,950 square miles [23,200 square kilometres]), Djibouti is bordered by Eritrea to the north, Ethiopia to the west and southwest, and Somalia to the south. The Gulf of Tadjoura, which opens into the Gulf of Aden, bifurcates the eastern half of the country and supplies much of its 230 miles (370 kilometres) of coastline. The capital, Djibouti city, is built on coral reefs jutting into the southern entrance of the gulf; other major towns are Obock, Tadjoura, Ali Sabieh, and Dikhil.

The nation's Lilliputian aspect belies its regional and geopolitical importance. The capital is the site of a modern deepwater port that serves Indian Ocean and Red Sea traffic and hosts a French naval base. Djibouti city is also the railhead for the only line serving Addis Ababa, the capital of Ethiopia.

The port of Djibouti

PHYSICAL AND HUMAN GEOGRAPHY

The land. The landscape of Djibouti is varied and extreme, ranging from rugged mountains in the north to a series of low desert plains separated by parallel plateaus in the west and south. Its highest peak is Mount Mousa at 6,768 feet (2,063 metres); the lowest point, which is also the lowest in Africa, is the saline Lake Assal, 515 feet (157 metres) below sea level. Located at the convergence of the African and Arabian tectonic plates, the territory is geologically active. Slight tremors are frequent, and much of the terrain is littered with basalt from past volcanic activity.

Rainfall is rare, and vegetation is minimal. There are no regularly flowing surface watercourses in the republic. Cool-season (October to April) daily maximum temperatures at Djibouti city average 87° F (31° C); in the hot months 99° F (37° C) is the average daily maximum. Temperatures increase and humidity drops in midsummer as the arid khamsin wind blows off the inland desert.

The country's wildlife includes antelopes, gazelles, hyenas, jackals, and ostriches. Offshore, Djibouti's waters teem with many species of marine life, including tuna, barracuda, and grouper.

Djibouti is virtually a city-state, since about two-thirds of the population lives in or near the capital. Outlying towns are small trading centres that experience periodic population increases as camel caravans and sheep and goat herders encamp.

The people. *Ethnic composition.* Based on linguistic criteria, the two largest ethnic groups are the Somali and the Afar. Both groups adhere at least nominally to the Sunnite branch of Islām and speak related, but not mutually intelligible, eastern Cushitic languages.

The Afar (Denakil, or Danakil) speak a language that forms a dialect continuum with Saho. Saho-Afar is usually classified as an Eastern Cushitic language of the Afro-Asiatic language phylum. The Afar live in the sparsely populated areas to the west and north of the Gulf of Tadjoura. This region includes parts of several former as well as extant Afar sultanates. The sultans' roles are now largely ceremonial, and the social divisions within the traditional Afar hierarchy are of diminished importance.

The Somali, who also speak an Eastern Cushitic language, are concentrated in the capital and the southeastern quarter of the country. Their social identity is determined by clan-family membership. More than half of the Somali belong to the Issa, whose numbers exceed those of the Afar; the remaining Somali are predominately members of the Gadabourisi and Issaq clans.

Djibouti city is home to a long-established community of Yemeni Arabs and houses a sizable contingent of French technical advisers and military personnel. In recent decades these groups have been joined by small but significant numbers of ethnic Ethiopians as well as Greek and Italian expatriates.

Foreign residents

Language. The republic recognizes two official languages: French and Arabic. However, Somali is the most widely spoken language. The use of Afar is mostly restricted to Afar areas.

Demographic trends. Djibouti is the most urbanized country in sub-Saharan Africa, with some four-fifths of the population classified as urban. The annual rate of population increase is higher than the world average but has dropped significantly since the 1980s. More than half of the population is under the age of 20, and the average life expectancy is roughly 50 years.

Both the Afar and the Somali maintain ties with relatives living in neighbouring Eritrea, Ethiopia, and Somalia. Since independence, many newcomers from rural areas and regions beyond the national frontier have migrated to live with family members in Djibouti city. Drought and political conflicts in the Horn also have created large refugee movements into the republic.

The economy. Djibouti has few natural resources and extensive unemployment. Efforts to exploit geothermal energy are under way, but without substantial results. Salt was commercially exploited for export until the 1950s; today, surface deposits are collected and marketed through the informal sector of the economy. In rural areas, nomadic pastoralism is a way of life. Sheep and goats are



Men quarrying salt at Lake Assal, Djibouti.
Victor Englebert

raised for milk, meat, and skins, while camels are used for transport caravans. Agriculture is confined to a few wadis, which produce small yields of vegetables (mostly tomatoes) and dates. The fishing industry is still in the early stages of development. More than 90 percent of the country's food requirements is imported, mainly from France, Kenya, and Ethiopia.

Much of the country's economic potential lies in the transport and service sectors. An international airport is located at Ambouli. The port of Djibouti is a free-trade zone with modern container and refrigeration facilities and a rail link to Ethiopia. International telecommunications services are some of the best in sub-Saharan Africa. The capital has attracted several large commercial banks and provides a thriving entertainment industry necessary to a port city. There is also much unrecorded transshipment, via camels, dhows, and trucks, to bordering countries.

Major public works projects have been funded through foreign aid, and the government actively coordinates donors' efforts. In 1988 a paved road linking Tadjoura and the north with the capital was completed. The improvement of housing and the urban infrastructure continues.

Administration and social conditions. *Government.* Nine constitutional articles were adopted in February 1981. These provide for the election of a president by universal suffrage for a six-year term (renewable once), a 65-member National Assembly elected for a five-year term, and a Council of Ministers headed by the prime minister.

The ruling party

A single-party system, consisting of the Popular Assembly for Progress (Rassemblement Populaire pour le Progrès; RPP), was instituted by constitutional amendment in October 1981. Deputies to the National Assembly must be elected from a list supplied by the RPP; abstention from voting is the only legal form of opposition.

The judicial system recognizes several codes: French-based civil law, Islāmic law, and customary means of arbitration employed by the local populations.

The Djiboutian armed forces are supported by the presence of several thousand French troops, including a unit of the French Foreign Legion.

Djibouti belongs to the United Nations, the Organization of African Unity, the Arab League, and the nonaligned movement. In 1986 Djibouti city became the headquarters of the Inter-Governmental Authority on Drought and Development (IGADD), which comprises six eastern African nations.

Education. The educational system, although free, is burdened by the needs of Djibouti's young population. For many, formal education ends with early childhood training at local Qur'ān schools. Primary schools are run by the state and by Roman Catholic clergy; advancement to the secondary level in the public system is limited by the size of state facilities. A small vocational training program is offered, but no postsecondary educational institutions exist. Less than one-fifth of the adult population is literate.

Health. Many Djiboutians live in poor housing with inadequate water and sanitation. The infant mortality rate is high due to diarrhea and dehydration. Tuberculosis is a major health problem. Djibouti city has a hospital and several primary care clinics, and local dispensaries serve the rural areas.

Cultural life. Djibouti's only television and radio station, which broadcasts in French, Arabic, Afar, and Somali, is state-run, as is the weekly French-language newspaper, *La Nation.* The government sponsors several organizations dedicated to the preservation of traditional culture and dance.

In 1984 Djibouti entered the Olympics for the first time; since then its marathon runners have commanded international attention.

Major holidays are Independence Day, June 27, and the festivals of the Muslim calendar.

For statistical data on the land and people of Djibouti, see the *Britannica World Data* section in the BRITANNICA BOOK OF THE YEAR.

HISTORY

On the eve of independence, Djibouti's viability as a sovereign state was questionable. However, fears that the Afar and the Issa Somali would become pawns in a struggle between the republic's rival neighbours, Ethiopia and Somalia, did not materialize. No Djiboutian political leader, either Afar or Somali, ever condoned unification with either of the larger states. Indeed, Djibouti established a peaceful international profile through a policy of strict neutrality in regional affairs. In keeping with friendship treaties with both Somalia and Ethiopia, the government refused to support armed groups opposing the neighbouring regimes, and it hosted negotiations between Somalia's and Ethiopia's leaders that resulted in a series of accords in 1988.

Balance between Afar and Somali

Djibouti's balanced posture in external relations was reflected in its internal politics. Hassan Gouled Aptidon, an Issa Somali, was elected to two consecutive terms as president in 1981 and 1987. Barkat Gourad Hamadou, an Afar, serving as prime minister since 1978, was reappointed in 1987. Power appeared to be shared, with ministry appointments following a formula designed to maintain ethnic balance.

In the first years of self-government, though, ethnic tensions were evident. By 1978 the state had experienced two cabinet crises and changes of prime minister. Those ousted were Afars accused of fomenting ethnic strife. Since the banning of opposition parties in 1981, ethnic conflict in the political arena has been for the most part minimal. However, Issa predominance in the civil service, the armed forces, and the RPP was only slightly masked, and occasional tremors of social unrest disturbed Djibouti's superficial calm.

Challenges to Djibouti's stability could not be reduced to traditional Afar and Issa enmity; signs of the serious problems facing the young nation were also to be found in the urban demography of its capital. On the outskirts of the city an expansive squatter community known as Balbala, which originally developed just beyond the barbed-wire boundary erected by the French colonial administration to prevent migration to the capital, tripled in size within a decade after independence. In 1987 it was officially incorporated into the city, with the promise of development of basic water and sanitary services. Its growth continued owing to a high birth rate, rural migration, and displacement of persons from the urban core.

Conditions in the densely populated "native" quarters of Djibouti city were only marginally better than in Balbala. Structures were limited to wood and corrugated iron by colonial, and later national, restrictions on the construction and location of permanent dwellings. Distinct ethnic enclaves were identifiable: the retail centre surrounding the main mosque (Hamoudi Mosque) and the former caravan terminus (Harbi Square), housing the Arab community; the neighbourhoods radiating beyond this area, settled by the Issaq, Gadaboursi, and Issa Somali; and the quarter known as Arhiba, built by the French to house the Afar dockworkers recruited from the north of the colony in the 1960s.

As the urban infrastructure was developed, and as government-subsidized housing was realized through international aid programs, conditions in the old districts of the city improved. Yet the needs remained immense, and progress was accompanied by perceptions of ethnic favouritism. Discontent was also fostered by a high cost of living, unemployment, and a widening gap in living conditions between the majority of the population and the new urban elite.

Influx of refugees

Finally, government efforts to assist the people in desolate rural areas were complicated by the presence of large numbers of Ethiopian and Somalian refugees. Under the auspices of the United Nations High Commissioner for Refugees, thousands of persons who were displaced by the Ogaden dispute of the late 1970s and the droughts of the early 1980s were repatriated. However, continued civil upheavals in Somalia precipitated more refugee movement into the republic. Thus, the chronic conflicts of the troubled Horn of Africa encumbered the realization of Djibouti's national goals of unity, equality, and peace.

(C.C.C.)

For later developments in the history of Djibouti, see the BRITANNICA BOOK OF THE YEAR.

Eritrea

Eritrea (Tigrinya: Ertra) is a small country of the Horn of Africa, located on the Red Sea. Its 600 miles (1,000 kilometres) of coastline extend from Cape Kasar, in the north, to the Strait of Mandeb, separating the Red Sea from the Gulf of Aden in the south. It is bounded on the northwest by The Sudan, on the south by Ethiopia, and on the southeast by Djibouti. Total land area (including islands off the coast) is 45,300 square miles (117,400 square kilometres). Eritrea's capital and largest city is Asmera.

Eritrea's coastal location has long been important in its history and culture—a fact reflected in its name, which is an Italianized version of Mare Erythraeum, Latin for "Red Sea." The Red Sea was the route along which Christianity and Islām reached the area and took firm hold among the people, and it was an important trade route that such powers as Turkey, Egypt, and Italy hoped to dominate by seizing control of ports on the Eritrean coast. Those ports promised access to the gold, coffee, and slaves sold by traders in the Ethiopian highlands to the south, and in the second half of the 20th century Ethiopia became the power from which the Eritrean people had to free themselves in order to create their own state. In 1993, after a war of independence that lasted nearly three decades, Eritrea became a sovereign country. During the long struggle, the people of Eritrea managed to forge a common national consciousness, but, with peace established, they now face the task of overcoming their ethnic and religious differences in order to raise the country from a poverty made worse by years of drought, neglect, and war.

PHYSICAL AND HUMAN GEOGRAPHY

The land. *Relief.* Eritrea's land is highly variegated. Running on a north-south axis through the middle of the country are the central highlands, a narrow strip of country some 6,500 feet (2,000 metres) above sea level that represents the northern reaches of the Ethiopian Plateau. Geologically, this plateau consists of a foundation of crystalline rock (*e.g.*, granite, gneiss, micaschist) that is overlain by sedimentary rock (limestone and sandstone) and then capped by basalt (rock of volcanic origin). The upper layers have been highly dissected by deep gorges and river channels, forming small steep-sided, flat-topped tablelands known as *ambas*. The highest point in the plateau is Mount Soira, at 9,885 feet (3,013 metres).

The central highlands

In the north of Eritrea the highlands narrow and then end in a system of hills, where erosion has cut down to the basement rock. To the east the plateau drops abruptly into a coastal plain. North of the Gulf of Zula, the plain is only 10 to 50 miles wide, but to the south it widens to include the Denakil Plain. This barren region contains a depression known as the Kobar Sink (more than 300 feet below sea level), the northern end of which extends into Eritrea. The coastal plain and the Denakil Plain are part of the East African Rift System and are sharply delimited on the west by the eastern escarpment of the plateau, which, although deeply eroded, presents a formidable obstacle to travelers from the coast.

The western flank of the central highlands is a broken and undulating plain that slopes gradually toward the border with The Sudan. It lies at an average elevation of 1,500 feet. The vegetation is mostly savanna, consisting of scattered trees, shrubs, and seasonal grasses.

Off the coast in the Red Sea is the Dahlak Archipelago, a group of more than 100 small coral and reef-fringed islands. Only a few of these islands have a permanent population.

Drainage. The Eritrean highlands are drained by four major rivers and numerous streams. Two of the rivers, the Gash and the Tekeze, flow westward into The Sudan. The Tekeze River (also known as the Satit) is a major tributary of the Atbara River, which eventually joins the Nile. The Gash River reaches the Atbara only during flood season. As it crosses the western lowlands, the Tekeze forms part of Eritrea's border with Ethiopia, while the upper course of the Gash, known as the Mereb River, forms the border on the plateau.

The other two major rivers that drain the highlands of Eritrea are the Barka and the Anseba. Both of these rivers flow northward into a marshy area on the eastern coast of The Sudan and do not reach the Red Sea. Several seasonal streams that flow eastward from the plateau reach the sea on the Eritrean coast.

Climate. Eritrea has a wide variety of climatic conditions, produced mainly by differences in altitude. The effects of elevation are seen most clearly in the wide range of temperatures experienced throughout the country. On the coast, Mitsiwa has one of the highest averages in the world (86° F, or 30° C), while Asmera, only 40 miles away yet more than 7,500 feet higher on the plateau, averages 62° F (17° C).

Mean annual rainfall on the plateau is 16 to 20 inches (400 to 500 millimetres), while on the western plain it is less than 16 inches. In both the highlands and the western lowlands, rainfall comes in summer, carried on a southwesterly airstream that decreases in amount of precipitation and length of rainy season as it proceeds toward the northeastern extremes of the plateau. The eastern edges of the plateau and, to a lesser extent, the coastal fringes receive much smaller quantities of rain from a northeasterly airstream that arrives in winter and spring. The interior regions of the Denakil Plain are practically rainless.

Settlement patterns. The environment is a determining factor in the distribution of Eritrea's population. Although

Mike Goldwater—Network/Matrix



The stable and walled courtyard of a farmer's house on a plateau near Mes-hal, a village south of Asmera, Eritrea.

Highland agriculturalists and lowland pastoralists

the plateau represents only one-quarter of the total land area, it is home to approximately one-half of the population, most of them sedentary agriculturalists. The lowlands on the east and west support a population mainly of pastoralists, although most of them also cultivate crops when and where weather conditions permit. As a rule, pastoralists follow various patterns of movement set by the seasons. Only the Rashaida, a small group in the northern hills, is truly nomadic.

Under Italian colonial rule from 1889 to 1941, Eritrea's urban sector flourished with the establishment of Asmera as the capital city, Aseb as a new port on the Red Sea, and a host of smaller towns on the plateau. In addition, Mitsiwa, an old and cosmopolitan port with strong links to Arabia, was expanded considerably. By the end of the colonial period, Eritrea had by far the highest urbanization rate in the Horn of Africa—approximately 15 percent—although a large part of the urban population was Italian nationals who eventually left the country. Subsequently, a population drift from the countryside to the towns was offset by emigration of Eritreans abroad, so that at the time of independence in 1993 the relative size of the urban sector remained unchanged.

The people. *Language groups.* Eritrea's population consists of several ethnic groups, each with its own language and cultural tradition. The Eritrean highlands are an extension of the Ethiopian Plateau to the south, and the bulk of the peasantry on the plateau belong to the Tigray, a group that also occupies the adjacent Ethiopian province of Tigray. The Tigrayan language, called Tigrinya, is spoken on both sides of the border and is the speech of nearly one-half of all Eritreans.

Inhabiting the northernmost part of the Eritrean plateau, as well as lowlands to the east and west, are people who speak the other major Eritrean language—Tigre. Tigre and Tigrinya are written in the same script and are descended from the same mother tongue (the ancient Semitic language of Ge'ez), but they are mutually unintelligible.

Also occupying the northern plateau are Bilin speakers, whose language belongs to the Cushitic family. The Rashaida are a group of Arabic-speaking nomads who traverse the northern hills.

On the southern part of the coastal region live Afar nomads, whose relatives live across the borders in Djibouti and Ethiopia; they are also called the Denakil, after the region that they inhabit. The coastal strip south of Mitsiwa, as well as the eastern flanks of the plateau, are occupied by Saho pastoralists. In the western plain, the dominant people are pastoralists of the Beja family, whose kin live across the border in The Sudan. Two small Nilotic groups, the Kunama and the Nara, also live in the west.

Religions. Historically, religion has been a prominent symbol of ethnic identity in the Horn of Africa. Christianity was established in the 4th century AD on the coast and appeared soon afterward in the plateau, where it was embraced by the Ethiopian highlanders. The Monophysite creed of the Ethiopian Orthodox church remains the faith of about half of the population of Eritrea, including nearly all the Tigray. Following the rise of Islām in Arabia, Muslim power flowed over the Red Sea coast, forcing the Ethiopians to retreat deep into their mountain fastness. Islām displaced other creeds in the lowlands of the Horn, and it remains the faith of nearly all the people inhabiting the eastern coast and the western plain of Eritrea, as well as the northernmost part of the plateau. Thus, while Islām claims nearly all pastoralists, Christianity is dominant among the peasant cultivators. (Muslims are significantly represented also in all towns of Eritrea, where they are prominent in trade.) In the perennial competition between cultivators and pastoralists over land, water, control of trade, and access to ports, religion has played an ideological role, and it remains a potent political force.

Christian highlanders and Muslim lowlanders

During the colonial period, Catholic and Protestant European missionaries introduced their own version of Christianity into Eritrea. They had considerable success among the small Kunama group, and they also attracted a few townspeople with the offer of modern education.

The economy. *Natural resources.* Salt mining, based on deposits in the Kobar Sink, is a traditional activity in Eritrea. Deposits of gold, copper, potash, and iron have been exploited at times in a minor way, and numerous other minerals have been identified, including zinc, feldspar, gypsum, asbestos, mica, and sulfur.

The area of cultivation is limited by climate and the uneven surface of the plateau, so that, of the 8 million acres (3.2 million hectares) of land considered cultivable, only 5 percent is being worked. There is room for expansion, however, especially if the country's considerable water resources are harnessed for irrigation.

In normal times, livestock is a valuable resource, and it has the potential to play a role in Eritrea's foreign trade. During the long war of independence, however, livestock was severely depleted. The fishing potential of the Red Sea is another underutilized resource.

Soil erosion, an age-old process, is particularly severe on the plateau. Encouraged by the steady expansion of cultivation, it has left few wooded areas and has created a shortage of fuel. The proximity of the oil-rich Arabian basin has occasionally raised expectations of discovering petroleum in Eritrea, but intermittent exploration since the days of Italian rule has failed to produce results.

Agriculture. Agriculture is by far the most important sector of the country's economy, providing a livelihood for about 80 percent of the population and normally accounting for the bulk of Eritrea's exports. Peasant cultivation and traditional pastoralism are the main forms of agricultural activity. These are not mutually exclusive occupations, since most peasants also keep animals and most pastoralists cultivate grains when possible. Both peasants and pastoralists produce primarily for their own subsistence, and only small surpluses are available for trade.

The staple grain products are an indigenous cereal named teff (*Eragostis abyssinica*) as well as corn (maize), wheat, barley, sorghum, and millet. Vegetables and fruit also are produced. Under Italian rule, modern irrigated plantations produced vegetables, fruit, cotton, sisal, bananas, tobacco, and coffee for the growing urban markets. This sector continued to operate under Ethiopian rule until it was disrupted by the long period of warfare.

Industry. A generation of war also damaged Eritrea's modest manufacturing sector, which also appeared during the Italian colonial period and provided many Eritrean workers with skills that enabled them later to find work abroad. Industry was based largely on the processing of agricultural products. Asmera was the main industrial centre, concerned with food products, beer, tobacco, textiles, and leather.

Red Sea ports

A petroleum refinery in the Red Sea port of Aseb, built by the Soviet Union for Ethiopia, is the prime industrial enterprise in Eritrea. Aseb also has a salt works, and there are a salt works and a cement works near the port of Mitsiwa.

Trade and transportation. Aseb and Mitsiwa have long been major ports of entry to Ethiopia, and that country, now landlocked, still has guarantees of access to the port facilities at Aseb. As a result, the bulk of Eritrea's trade is in the transit of goods to Ethiopia. A paved road links Aseb with Addis Ababa, and another paved road begins in Mitsiwa, climbs the plateau to Asmera, and continues south to the Ethiopian town of Adigrat. A railway was built by the Italians from Mitsiwa to Asmera, Keren, and Akordat, but it was rendered useless by the war of independence.

There are an international airport in Asmera and major airfields in Aseb and Mitsiwa.

Administration and social conditions. *Government.* After liberation from Ethiopia in May 1991, Eritrea was ruled by a provisional government that consisted essentially of the central committee of the Eritrean People's Liberation Front (EPLF). On May 19, 1993, shortly after a national referendum, this body proclaimed the Transitional Government of Eritrea, which was to rule for four years until the promulgation of a constitution and the election of a permanent government. The transitional government's legislative body, called the National Assembly, consisted of the original 30-member central committee of the EPLF augmented by 60 new members. At least 20 seats were to be reserved for women.

The National Assembly sets the policies of the government and elects the president. The president is assisted in implementing the government's policies by a State Council, containing cabinet ministers and the governors of Eritrea's provinces. In order to discourage ethnic rivalry, seats on the State Council are divided equally between Muslims and Christians, and political parties based on language or religion are banned.

Health and education. Chronic drought and decades of war have taken a toll on the health of Eritreans. The mortality rate at birth is 15 percent, and almost half of all infants die during their first year. The average life expectancy is about 50 years.

Only about 20 percent of Eritreans are literate, though the new government is intent on expanding education. Children are taught in their native languages, and in the higher grades they also are taught foreign languages, especially Arabic and English. There is a university in Asmera.

For statistical data on the land and people of Eritrea, see the *Britannica World Data* section in the BRITANNICA BOOK OF THE YEAR.

HISTORY

Precolonial Eritrea. *Rule from the highlands.* Beginning about 1000 BC, Semitic peoples from the South Arabian kingdom of Saba' (or Sheba) migrated across the Red Sea and absorbed the Cushitic inhabitants of the Eritrean coast and adjacent highlands. These Semitic invaders, possessing a well-developed culture, established the kingdom of Aksum, which, by the end of the 4th century AD, ruled the northern stretches of the Ethiopian Plateau and the eastern lowlands. An important trade route led from the port of Adulis, near modern Zula, to the city of Aksum, the capital, located in what is now the Ethiopian province of Tigray.

After extending its power at times as far afield as modern Egypt and Yemen, Aksum began to decline into obscurity in the 6th century AD. Beginning in the 12th century, however, the Ethiopian Zagwe and Solomonid dynasties held sway to a fluctuating extent over the entire plateau and the Red Sea coast. Eritrea's central highlands, known as the *mereb melash* ("land beyond the Mereb River"), were the northern frontier region of the Ethiopian kingdoms and were ruled by a governor titled *bahr negash* ("lord of the sea"). The control exercised by the crown over this region was never firm, and it became even more tenuous as the centre of Ethiopian power moved steadily southward to Gonder and Shewa. Highland Eritrea became a vassal fiefdom of the lords of Tigray, who were seldom on good terms with the dominant Amhara branch of the Ethiopian family.

The "land beyond the Mereb"

Contesting for the coastlands. Off the plateau, the pastoralist peoples in the west and north knew no foreign master until the early 19th century, when the Egyptians invaded the Sudan and raided deep into the Eritrean lowlands. The Red Sea coast, owing to its strategic and commercial importance, was contested by many powers. In the 16th century the Turks occupied the Dahlak Archipelago and then Mitsiwa, where they maintained, with occasional interruption, a garrison for more than three centuries. Also in the 16th century, Eritrea as well as Ethiopia were affected by the invasions of Aḥmad Grāñ, the Muslim leader of the sultanate of Adal. After the expulsion of Aḥmad's forces, the Turks temporarily occupied even more of Eritrea's coastal area. In 1865 the Egyptians obtained Mitsiwa from the Ottoman Porte. From there they pushed inland to the plateau, until in 1875 an Egyptian force that reached the Mereb River was annihilated by Ethiopian forces.

Meanwhile, the opening of the Suez Canal in 1869 had made the Red Sea a scene of rivalry among the world's most powerful states. Between 1869 and 1880 the Italian Rubattino Navigation company purchased from the local Afar sultan stretches of the Red Sea coast adjoining the village of Aseb. In 1882 these acquisitions were transferred to the Italian state, and in 1885 Italian troops landed at Mitsiwa, Aseb, and other locations. There was no resistance by the Egyptians at Mitsiwa, and protests made by the Turks and Ethiopians were ignored. Italian forces then systematically spread out from Mitsiwa toward the highlands. This expansion onto the plateau was initially opposed by Emperor Yohannes IV, the only Tigray to wear the Ethiopian crown in modern times, but Yohannes's successor, Menilek II, in return for weapons that he needed to fight possible rivals, acquiesced to Italian occupation of the region north of the Mereb. In the Treaty of Wichale, signed on May 2, 1889, Menilek recognized "Italian possessions in the Red Sea," and on Jan. 1, 1890, the Italian colony of Eritrea was officially proclaimed. From here, the Italians launched several incursions into Ethiopia, only to be decisively defeated by Menilek's army at the Battle of Adwa on March 1, 1896. Menilek did not pursue the defeated enemy across the Mereb. Soon afterward, he signed the Treaty of Addis Ababa, obtaining Italian recognition of Ethiopia's sovereignty in return for his recognition of Italian rule over Eritrea.

Colonial Eritrea. *Ruled by Italy.* In precolonial times there were no towns on the Eritrean plateau, urban centres being limited to the Red Sea coast. Under Italian rule, however, Eritrea's urban sector flourished. Tens of thousands of Italians arrived, bringing with them modern skills and a new lifestyle. Asmera grew into a charming city in Mediterranean style, the port of Mitsiwa was modernized and the port of Aseb improved, and a number of smaller towns appeared on the plateau. Road and rail construction linked the various regions of the colony, and a modest manufacturing sector also appeared, so that Eritreans acquired industrial skills.

Economic growth and modernization

At the same time, a sizable portion of Eritrea's best agricultural land was reserved for Italian farmers (although only a few actually settled on the land), and a small plantation sector was established to grow produce for the urban market. Eritrea's population grew rapidly during this period. Combined with the appropriation of land for Italian use, population growth created a shortage of land for the peasantry. This in turn stimulated a drift to the cities, which further expanded the urban population and produced an Eritrean working class.

Still, Eritrea had no valuable resources for exploitation and was not a wealth-producing colony for Italy. In fact, the colony was subsidized by the Italians, an extraneous factor that gave the local economy an artificial glow. Investment in education for Eritreans was negligible. There were very few schools for them, and even these were limited to the primary level. Also, Eritreans were not employed in the colonial service except as labourers and soldiers. As preparations for the invasion of Ethiopia got under way in the mid-1930s, several thousand Eritreans were recruited to serve in the invading army.

From Italian to Ethiopian rule. The invasion and occupation of Ethiopia beginning in 1935 marked the last chapter in Italian colonial history—a chapter that came to an end with the eviction of Italy from the Horn of Africa by the British in 1941. The following decade, during which Eritrea remained under British administration, was a period of intense political and diplomatic activity that shaped the future of Eritrea. Landlocked Ethiopia, coveting Eritrea's two seaports, launched an early campaign to annex the former colony, claiming that it had always been part of Ethiopia's domain. Lobbying of the Allied powers was carried out, and within Eritrea support for annexation was mobilized on the basis of religious loyalty by utilizing the services of Ethiopian Orthodox clergy. In order to promote the union of Eritrea with Ethiopia, a Unionist Party was formed in 1946; it was financed and guided from Addis Ababa.

Eritrea's Muslims had every good reason to oppose union with Ethiopia, where Christianity was the official religion and Muslims suffered discrimination in many areas of life. In order to counter Christian mobilization for union, a Muslim League was founded in 1947 to campaign for Eritrean independence. Thus, although there were some Christians who favoured independence and a few Muslims who were favourable to union with Ethiopia, the political division was drawn largely along sectarian lines.

Federation with Ethiopia. *Adoption of the federal scheme.* In 1950 the United Nations (UN), under the prompting of the United States, resolved to join Eritrea

to Ethiopia within two years in a federation that would provide the former colony with autonomy under its own constitution and elected government. Elections to a new Eritrean Assembly in 1952 gave the Unionist Party the largest number of seats—but not a majority, so it formed a government in coalition with a Muslim faction. The Eritrean constitution, prepared by the UN in consultation with Emperor Haile Selassie I of Ethiopia, was adopted by the Eritrean Assembly on July 10, 1952, and ratified by Haile Selassie on August 11. The act of federation was ratified by the emperor on September 11, and British authorities officially relinquished control on September 15.

Failure of the federal scheme. The federal scheme was short-lived, mainly because the government in Addis Ababa was unwilling to abide by its provisions. First, the Eritrean constitution sought to establish an equilibrium between ethnic and religious groups. It made Tigrinya and Arabic the official languages of Eritrea, and it allowed local communities to choose the language of education for their children. In the spirit of the constitution, the practice evolved of ensuring parity between Christians and Muslims in appointment to state office. This delicate balance was destroyed by Ethiopian interference, and Muslims were the initial losers, as Arabic was eliminated from state education and Muslims were squeezed out of public employment.

Suppression of Eritrean nationalism

Furthermore, the Ethiopians were anxious to eliminate any traces of separatism in Eritrea, and to that end they harrassed the leaders of the independence movement until many of them fled abroad. With the collaboration of their Unionist allies and in express violation of the constitution, they also suppressed all attempts to form autonomous Eritrean organizations. Political parties were banned in 1955, trade unions were banned in 1958, and in 1959 the name Eritrean Government was changed to "Eritrean Administration" and Ethiopian law was imposed. Eventually, even Ethiopia's Eritrean allies were alienated by crude intervention in the running of the Eritrean administration, financial disputes between Asmera and Addis Ababa, and mounting pressure on the Eritreans to renounce autonomy. The federation was already dead when, on Nov. 14, 1962, the Ethiopian parliament and Eritrean Assembly voted unanimously for the abolition of Eritrea's federal status, making Eritrea a simple province of the Ethiopian empire. Soon afterward, Tigrinya was banned in education; it was replaced by Amharic, the official language of Ethiopia.

The war of independence. *Beginning of armed revolt.* Muslims had been the first to suffer from Ethiopia's intervention in Eritrea, and it was they who formed the first opposition movement. In 1960, leaders of the defunct independence movement who were then living in exile announced the formation of the Eritrean Liberation Front (ELF). The founders, all Muslims, were led by Idris Mohammed Adam, a leading political figure in Eritrea in the 1940s. By the mid-1960s the ELF was able to field a small guerrilla force in the western plain of Eritrea, and thus began a war that was to last nearly three decades. In the early years, the ELF drew support from Muslim communities in the western and eastern lowlands as well as the northern hills. It also sought support from The Sudan, Syria, Iran, and other Islāmic states, and used Arabic as its official language. Ethiopian authorities portrayed the movement as an Arab tool and sought to rally Eritrean Christians to oppose it. Deteriorating economic and political conditions in Eritrea, however, combined to produce the opposite result.

During the 1930s and '40s the Eritrean economy had been stimulated by Italian colonial activity and by the special conditions created by World War II. After the war the local economy deflated, and it remained stagnant during the entire period of federation with Ethiopia. Many thousands of Eritreans were forced to emigrate to Ethiopia and the Middle East in search of employment. The suppression of the nascent trade-union movement further embittered this class, and many Eritrean workers—Muslims and Christians alike—rallied to the nationalist movement. In addition, the banning of Tigrinya in state education helped to turn an entire generation of Eritrean Christian students toward nationalism. Christians began to join the ELF in significant numbers at the end of the 1960s. Among them were students who had become politically radicalized in the Ethiopian student movement, which itself became a centre of opposition to the regime of Haile Selassie in the 1960s and '70s.

Highlanders join the rebellion

The spreading of revolution. The ELF was now able to extend its operations to the central highlands of Eritrea—the home of the Tigray. However, the arrival of the radical students coincided with the emergence of a serious rift between the leadership of the ELF, which was permanently resident in Cairo, and the rank and file, which remained in the field. The newcomers joined the opposition to the leadership, and in 1972 several groups that had defected from the ELF joined forces to form the Eritrean Liberation Front–People's Liberation Forces (ELF–PLF). For several years the two rival organizations fought each other as well as the Ethiopians. After a series of splits and mergers, the ELF–PLF came under the control of former students, among whom Christians predominated, and was renamed the Eritrean People's Liberation Front (EPLF), a Marxist and secular organization.

The EPLF had made its presence felt by 1974, when the imperial regime in Ethiopia collapsed. While a power struggle for the succession was waged in Addis Ababa, the two Eritrean fronts liberated most of Eritrea. By 1977 the nationalist revolution seemed on the verge of victory. It was not to be. A military dictatorship emerged in Addis Ababa—also espousing Marxism—and, armed and assisted by the Soviet bloc, the new Ethiopian regime was able to regain most of Eritrea in 1978. Warfare on a scale unprecedented in sub-Saharan Africa raged for the next decade. The Ethiopians made enormous efforts with massive land attacks and heavy weaponry, but they had no success against the small and lightly armed guerrilla forces.

The violence of war and indiscriminate oppression in their homeland turned most Eritreans against Ethiopia, thereby producing a steady stream of young recruits for the nationalist movement. Throughout the 1980s the fighting was carried out by the EPLF, which by 1981 had succeeded in eliminating the ELF and had emerged as the unchallenged champion of Eritrean nationalism. In the latter part of the decade, the Soviet Union terminated its military aid to Ethiopia. Unable to find another patron and faced with armed rebellion in other parts of the country, the regime in Addis Ababa began to falter. The final act occurred in 1991, when a rebel military offensive, led by the Tigray People's Liberation Front, swept toward the capital. The Ethiopian army disintegrated, and in May the EPLF assumed complete control of Eritrea.

Independence

Three decades of war had produced among Eritreans a sense of unity and solidarity that they had not known before. Indeed, an entire generation had come of age during the struggle for independence, which was now to become a reality. The new regime in Ethiopia supported Eritrea's independence, so that a separation was effected amicably. In a referendum held two years after liberation, on April 23–25, 1993, the overwhelming majority of Eritreans voted for independence. On May 21, Isaias Afwerki, the secretary-general of the EPLF, was made president of a transitional government, and on May 24 he proclaimed Eritrea officially independent. (G.C.L./Jo.Ms.)

Following independence, Eritrea enjoyed a thriving economy but, with the noteworthy exception of Ethiopia, maintained poor relations with neighbouring nations. Tension with The Sudan throughout the 1990s centred on mutual allegations that each had attempted to destabilize the other, and in late 1995 Eritrea engaged in a brief but violent conflict with Yemen over the Ḥanīsh Islands, a group of Red Sea islands claimed by both countries. Postindependence relations with Ethiopia, heretofore warm and supportive, deteriorated rapidly in 1998 and exploded into violence over the disputed hamlet of Badme. Following two years of bloodshed, a peace was negotiated in December 2000, but Eritrea's earlier economic and political progress had been shattered. Amid economic distress, loss of life, and a new flood of displaced persons, the first voices of discontent with government leadership were raised, and calls were made to promulgate the nation's constitution, which had been ratified in 1997. (Ed.)

For later developments in the history of Eritrea, see the BRITANNICA BOOK OF THE YEAR.

Ethiopia

Ethiopia (Amharic: Ityop'iya) is a landlocked country in the Horn of Africa. It shares frontiers with Eritrea to the north, Djibouti to the northeast, Somalia to the east, Kenya to the south, and The Sudan to the west and northwest. Its total area is 437,794 square miles (1,133,882 square kilometres). Lying completely within the tropical latitudes, the country is relatively compact, with similar north-south and east-west dimensions. The capital is Addis Ababa ("New Flower"), located almost at the centre of the country.

Ethiopia is one of the oldest countries in the world. Its territorial extent has varied over the millennia of its existence. In ancient times it remained centred around Aksum, an imperial capital located in the northern part of the modern state, about 100 miles (160 kilometres) from the Red Sea coast. The present territory was consolidated during the 19th and 20th centuries as European powers encroached into Ethiopia's historical domain. Ethiopia became prominent in modern world affairs first in 1896, when it defeated colonial Italy in the Battle of Adwa, and again in 1935–36, when it was invaded and occupied by fascist Italy. Liberation during World War II by the Allied powers set the stage for Ethiopia to play a more prominent role in world affairs. Ethiopia was among the first independent nations to sign the Charter of the United Nations, and it gave moral and material support to the decolonization of Africa and to the growth of Pan-African cooperation. These efforts culminated in the establishment of the Organization of African Unity and the United Nations Economic Commission for Africa, both of which have their headquarters in Addis Ababa.

Ethiopia's prominence in Africa and elsewhere faded during the 17-year (1974–91) rule of the Derg, a Marxist regime that brought the country to the verge of disaster with civil wars aggravated by famine and starvation. With yet another provisional government at the helm, its political future and economic prospects remain uncertain.

PHYSICAL AND HUMAN GEOGRAPHY

The land. *Geology.* Ethiopia's topography, one of the most rugged in Africa, is built on four geologic formations. Rocks of Precambrian origin (more than 540 million years in age) form the oldest basal complex of Ethiopia, as they do in most of Africa. The Precambrian layer is buried under more recent geologic formations—except in parts of northern, western, and southern Ethiopia, where there are exposed rock layers of granite and schist. Geologic processes of the Mesozoic Era (245 to 66.4 million years ago) contributed sedimentary layers of limestone and sandstone, most of which have been either eroded or covered by volcanic rocks. Younger sedimentary layers are found in northern Ethiopia and on the floors of the Rift Valley. Lava flows from the Tertiary and Quaternary periods (from 66.4 million years ago to the present) have formed basaltic layers that now cover two-thirds of Ethiopia's land surface with a thickness ranging from about 1,000 feet (300 metres) to almost 10,000 feet. The Rift Valley forms a spectacular graben (a massive tectonic trough) running right down the middle of the country from the northern frontier with Eritrea to the southern border with Kenya.

Relief. Although Ethiopia's complex relief defies easy classification, five topographic features are discernible. These are the Western Highlands, Western Lowlands, Eastern Highlands, Eastern Lowlands, and Rift Valley. The Western Highlands are the most extensive and rugged topographic component of Ethiopia. The most spectacular portion is the North Central massifs; these form the roof of Ethiopia, with elevations ranging from 15,157 feet (4,620 metres) for Mount Ras Dejen (or Dashen), the highest mountain in Ethiopia, to the Blue Nile and Tekeze river channels 10,000 feet below.

The Western Lowlands stretch north-south along the Sudanese border and include the lower valleys of the Blue Nile, Tekeze, and Baro rivers. With elevations of about 3,300 feet, these lowlands become too hot to attract dense settlement.

The Rift Valley is part of the larger East African Rift System. Hemmed in by the escarpments of the Western and Eastern Highlands, it has two distinct sections. The first part is in the northeast, where the valley floor widens into a funnel shape as it approaches the Red Sea and the Gulf of Aden. This is a relatively flat area interrupted only by occasional volcanic cones, some of which are active. The Denakil Plain, in which a depression known as the Kobar Sink drops as low as 380 feet below sea level, is found here. High temperatures and lack of moisture make the northeastern Rift Valley unattractive for settlement. The southwest section, on the other hand, is a narrow depression of much higher elevation. It contains Ethiopia's Lakes Region, an internal drainage basin of many small rivers that drain into Lakes Abaya, Abiyata, Awasa, Langano, Shala, Chamo, and Ziway. Together these lakes have more than 1,200 square miles (3,108 square kilometres) of water surface. The upper Rift Valley is one of the most productive and most settled parts of Ethiopia. The Rift Valley

The Eastern Highlands are much smaller in extent than the Western Highlands, but they offer equally impressive contrast in topography. The highest peaks are Mount Batu, at 14,127 feet, and Mount Chilalo, at 13,575 feet. The Eastern Lowlands resemble the long train of a bridal gown suddenly dipping from the narrow band of the Eastern Highlands and gently rolling for hundreds of miles to the Somalian border. Two important regions here are the Ogaden and the Hawd. The Shebele and Genale rivers cross the lowlands, moderating the desert ecology.

Drainage. Ethiopia has three principal drainage systems. The first and largest is the western system, which includes the watersheds of the Blue Nile (known as the Abay in Ethiopia), the Tekeze, and the Baro rivers. All three rivers flow west to the White Nile in The Sudan. The second system is the Rift Valley internal drainage system, composed of the Awash River, the Lakes Region, and the Omo River. The Awash flows northeast to the Denakil Plain before it dissipates into a series of swamps and Lake Abe at the border with Djibouti. The Lakes Region is a self-contained drainage basin, and the Omo flows south into Lake Rudolf, on the border with Kenya. The third system is that of the Shebele and Genale rivers. Both of these rivers originate in the Eastern Highlands and flow southeast toward Somalia and the Indian Ocean. Only the Genale (known as the Jubba in Somalia) makes it to the sea; the Shebele (in Somali, Shabeelle) disappears in sand just inside the coastline.

Soils. The soils of Ethiopia can be classified into five principal types. The first type is composed of euritic nitosols and andosols and is found on portions of the Western and Eastern Highlands. These soils are formed from volcanic material and, with proper management, have medium to high potential for rain-fed agriculture. The second group of soils, eutric cambisols and ferric and orthic luvisols, are found in the Simen plateau of the Western Highlands. They are highly weathered with a subsurface accumulation of clay and are characterized by low nutrient retention, surface crusting, and erosion hazards. With proper management, they are of medium agricultural potential. Fertile volcanic soils

The third group of soils is the dark clay soils found in the Western Lowlands and at the foothills of the Western Highlands. Composed of vertisols, they have medium to high potential for both food and agriculture but pose tillage problems because they harden when dry and become sticky when wet. Some of the rich coffee-growing regions of Ethiopia are found on these soils.

The fourth group is composed of yermosols, xerosols, and other saline soils that cover desert areas of the Eastern Lowlands and the Denakil Plain. Because of moisture deficiency and coarse texture, they lack potential for rain-fed agriculture. However, the wetter margins are excellent for livestock, and even the drier margins respond well to irrigation. The fifth soil group is lithosols found primarily in the Denakil Plain. Lack of moisture and shallow profile preclude cultivation of these soils.

Soil erosion is a serious problem in Ethiopia. Particularly in the northern provinces, which have been settled with sedentary agriculture for millennia, population density has caused major damage to the soil's physical base,

to its organic and chemical nutrients, and to the natural vegetation cover. Even on the cool plateaus, where good volcanic soils are found in abundance, crude means of cultivation have exposed the soils to heavy seasonal rain, causing extensive gully and sheet erosion.

Climate. Because Ethiopia is located in the tropical latitudes, its areas of lower elevation experience climatic conditions typical of tropical savanna or desert. However, relief plays a significant role in moderating temperature, so that higher elevations experience weather typical of temperate zones. Thus, average annual temperatures in the highlands are about 61° F (16° C), while the lowlands average about 82° F (28° C).

The three seasons

There are three seasons in Ethiopia. From September to February is the long dry season known as the *bega;* this is followed by a short rainy season, the *belg,* in March and April. May is a hot and dry month preceding the long rainy season (*krem*t) in June, July, and August. The coldest temperatures generally occur in December or January (*bega*) and the hottest in March, April, or May (*belg*). However, in many localities July has the coldest temperatures because of the moderating influence of rainfall.

Ethiopia can be divided into four rainfall regimes. Rain falls year-round in the southern portions of the Western Highlands, where annual precipitation may reach 80 inches (2,000 millimetres). Summer rainfall is received by the Eastern Highlands and by the northern portion of the Western Highlands; annual precipitation there may amount to 55 inches. The Eastern Lowlands get rain twice a year, in April–May and October–November, with two dry periods in between. Total annual precipitation varies between 20 and 40 inches. The driest of all regions is the Denakil Plain, which receives less than 20 inches and sometimes none at all.

Plant and animal life. Ethiopia's natural vegetation is influenced by four biomes. The first is savanna, which, in wetter portions of the Western Highlands, consists of montane tropical vegetation with dense, luxuriant forests and rich undergrowth. Drier sections of savanna found at lower elevations of the Western and Eastern Highlands contain tropical dry forests mixed with grassland. The second biome is mountain vegetation; comprising montane and temperate grasslands, this covers the higher altitudes of the Western and Eastern Highlands. The third biome, tropical thickets and wooded steppe, is found in the Rift Valley and Eastern Lowlands. The fourth biome is desert steppe vegetation, which covers portions of the Denakil Plain.

Ethiopia has had a rich variety of wildlife that in some cases has been reduced to a few endangered remnants. Lions, leopards, elephants, giraffes, rhinoceroses, and wild buffalo are rarities, especially in northern Ethiopia. The Rift Valley, the Omo River valley, and the Western Lowlands contain remnants of big-game varieties. Smaller game varieties such as foxes, jackals, wild dogs, and hyenas are found abundantly throughout the country.

Uniquely Ethiopian and among the most endangered species are the *walia* ibex of the Simen Mountains, the mountain nyala (a kind of antelope), the Simien jackal, and the gelada monkey. They are found in the Western and Eastern Highlands in numbers ranging from a few hundred for the *walia* ibex to a few thousand for the others. More abundant varieties found in the lowlands include such antelopes as the oryx, greater kudu, and waterbuck, various types of monkeys including the black-and-white colobus (known as *guereza* in Ethiopia and hunted for its beautiful long-haired pelt), and varieties of wild pig. In order to protect remaining species, the government has set aside 20 national parks, game reserves, and sanctuaries covering a total area of 21,320 square miles—about 5 percent of the total area of Ethiopia.

Settlement patterns. With only about 12 percent of the population urbanized, most Ethiopians live in scattered rural communities. In order to reduce traveling distance, homesteads are generally scattered to be near farm plots. Buildings vary between circular and rectangular styles and are constructed of materials readily found within the environment. Roofs are mostly thatched, although some well-to-do farmers opt for corrugated steel tops.

Modern urban centres in Ethiopia include the national capital of Addis Ababa and such regional centres as Dire Dawa (in the east), Jima (south), Nekemte (west), Dese (north-central), Gonder (northwest), and Mekele (north). Addis Ababa, founded by Menilek II in 1886, brought an end to the custom of "roving capitals" practiced by earlier monarchs. After World War II, "Addis" obtained the lion's share of investments in industry, social services, and infrastructure, so that it became the most attractive place for young people to seek opportunity.

The people. Ethiopians are ethnically diverse, but it is not helpful to attempt to distinguish among peoples by physical criteria alone. The most important differences are cultural, particularly in language and religion.

Languages. Ethiopia is a mosaic of about 100 languages that can be classified into four groups—Semitic, Cushitic, Omotic, and Nilotic. The Semitic languages are spoken primarily in the northern and central parts of the country; they include Ge'ez, Tigrinya, Amharic, Gurage, and Hareri. Ge'ez, the ancient language of the Aksumite empire, is used today only for religious writings and worship in the Ethiopian Orthodox church. Tigrinya is native to the northern province of Tigray. Amharic is the national language and is native to the central and northwestern provinces. Gurage and Hareri are spoken by relatively few people in the south and east.

The dominant Semitic languages

The most important Cushitic languages are Oromo, Somali, and Afar. Oromo, together with Amharic and Tigrinya, is one of the three most-spoken languages in Ethiopia; it is native to the western, southwestern, southern, and eastern areas of the country. Somali is dominant among inhabitants of the Ogaden and Hawd, while Afar is most common in the Denakil Plain.

The Omotic languages, chief among which is Walaita, are not widespread, being spoken mostly in the densely populated areas of the extreme southwest. The Nilotic language group is native to the Western Lowlands, with Kunama speakers being dominant.

Religion. Christianity was introduced to Ethiopia in the 4th century, and the Ethiopian Orthodox church (called Tewahdo in Ethiopia) is one of the oldest Christian sects in the world. The church has long enjoyed a dominant role in the culture and politics of Ethiopia, counting more than half of all Ethiopians (including most of the Amhara and Tigray) among its adherents and having served as the official religion of the ruling elite until the demise of the monarchy in 1974. It also has served as the repository of Ethiopia's literary tradition and its visual arts. The core area of Christianity is in the highlands of northern Ethiopia, but its influence is felt in the entire country.

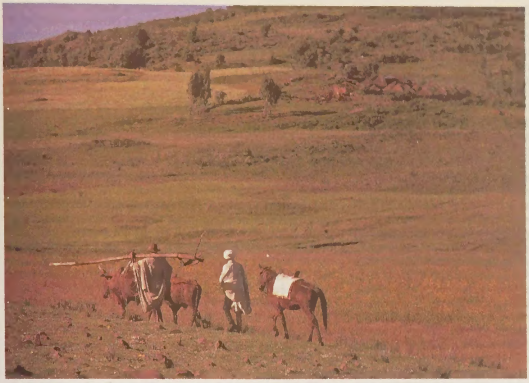
Islām was introduced in the 7th century and is now practiced by more than one-quarter of Ethiopians. It is most important in the outlying regions, particularly in the Eastern Lowlands, but there are local concentrations throughout the country. Traditionally, the status of Islām has been far from equal with that of Christianity. However, the emperor Haile Selassie gave audiences to Muslim leaders and made overtures in response to their concerns, and under the Derg even more was done to give at least symbolic parity to the two faiths. Nevertheless, the perception of Ethiopia as "an island of Christianity in a sea of Islām" has continued to prevail among both highland Ethiopians and foreigners. There are now some concerns among highlanders that fundamentalist Muslim movements in the region and in neighbouring countries may galvanize sentiments for a greater role of Islām in Ethiopia.

About one-tenth of Ethiopians are animists who worship a variety of African deities. They are primarily located in the Western Lowlands and speak a variety of Nilotic languages, such as Kunama.

Judaism has long been practiced in the vicinity of the ancient city of Gonder. Most of the Ethiopian Jews—who call themselves Beta Israel but also have been known as Falasha—have relocated to Israel.

The economy. Under Emperor Haile Selassie (reigned 1930–74), Ethiopia's economy enjoyed a modicum of free enterprise. The production and export of cash crops such as coffee were advanced, and import-substituting manu-

Free enterprise and state ownership



Farmer carrying plow, near Debre Markos, Western Highlands, Ethiopia.
Brian Seed from TSW—CLICK/Chicago

factures such as textiles and footwear were established. Especially after World War II, tourism, banking, insurance, and transport began to contribute more to the national economy. The Derg regime, which ruled from 1974 to 1991, nationalized all means of production, including land, housing, farms, and industry. Faced with uncertainties on their land rights, the smallholding subsistence farmers who form the backbone of Ethiopian agriculture became reluctant to risk producing surplus foods for market. Food shortages, already made serious by drought and civil war, worsened, and famine continued until the Derg finally collapsed. Under the present regime, which is essentially extracted from a rebel faction, land is still state-owned and is tenurable only by leasing from the government. In addition, an uncertain political climate has precluded significant internal or external investment in the country's economy. Ethiopia therefore remains among the poorest countries in the world.

Resources. Ethiopia's most promising resource is its agricultural land. Although soil erosion, overgrazing, and deforestation have seriously damaged the plateaus, nearly half the potentially cultivable land is still available for future use. Most of the reserve land is located in parts of the country that have favourable climatic conditions for intensive agriculture. In addition, Ethiopia is the richest country in Africa in number of livestock, including cattle. With better management of grazing lands and breeding, livestock raising has the potential to meet the demands of internal as well as export markets.

Ethiopia has many large rivers, but, with the exception of the Awash, they have yet to be exploited fully for hydroelectric power and irrigation. The role of minerals in Ethiopia's economy is also small. Only gold and platinum are of significance. However, there are potentials for copper, potash, lead, manganese, aluminum, chromium, cobalt, sulfur, and many others.

Agriculture. Agriculture contributes almost half of Ethiopia's gross domestic product (GDP). There are three types of agricultural activity. The first (and by far the most important) is the subsistence smallholder sector, which produces most of the staple grains such as teff, wheat, barley, and oats (on the cooler plateaus) and sorghum, corn (maize), and millet (in warmer areas) and pulses such as chickpeas, peas, beans, and lentils. Farm plots are very small, ranging from 3 to 6 acres (1.2 to 2.5 hectares).

The second type of agriculture is cash-cropping. Products include coffee, oilseeds, beeswax, qat (a stimulant leaf), and sugarcane. Coffee, which is native to Ethiopia, is the single most important export.

Subsistence livestock raising, the third agricultural activity, is important in the peripheral lowlands of Ethiopia. Large herds may be kept by a family as it migrates each season in search of grazing and water.

Fishing. Fish is not a major item in the diet of highland Ethiopians (except during Lent for Christians). Favourite species are harvested from freshwater lakes and rivers in the highlands and the Rift Valley. Most of the fish sold locally is produced by small operators whose scale of operation and technology is inadequate for export production.

Mining and quarrying. Compared to its potential, this sector contributes very little to the country's economy (less than 1 percent of its GDP). Gold is mined at Kibre Mengist in the south and platinum at Yubdo in the west. Also important are rock salt from the Denakil Plain and quarried building materials such as marble.

Modern manufacturing and cottage industry

Industry. Modern manufacturing contributes only about 7 percent to the GDP of Ethiopia. Products are primarily for domestic consumption. Among the most important are processed foods and beverages, textiles, tobacco, leather and footwear, and chemical products. Cottage industry and small enterprises are more important than industrial manufacturing in offering nonfarm employment and in producing a variety of consumer goods—for example, furniture, farming and construction implements, utensils, woven fabric, rugs, leathercrafts, footwear, jewelry, pottery, and baskets. Some of these products reach the tourist market. The size of the cottage industry is not known exactly, but government estimates put it at 4 percent of GDP.

Energy. Energy is derived primarily from firewood and charcoal. Ethiopia's long dependence on these sources has contributed to the depletion of its trees and to the erosion of its soil. Hydroelectricity, the most important source of power for industries and major cities, is generated at three stations on the Awash, two on the Blue Nile or its tributaries, and one on the Shebele. However, these represent only a small fraction of Ethiopia's potential.

A refinery at the Eritrean port of Aseb supplies Ethiopian motor vehicles and trains with gasoline and diesel fuel. Ethiopia itself does not extract petroleum.

Finance. Before the Derg's nationalization of banks and insurance companies, financial institutions were among the best-run and best-managed establishments in Ethiopia. The National Bank of Ethiopia is the country's central bank and is also responsible for regulatory functions. The Commercial Bank of Ethiopia is the largest commercial bank, with branches throughout the country. The Agricultural and Industrial Development Bank provides

loans primarily for agricultural and livestock development; it also directs small amounts toward investment in manufacturing.

Trade. Ethiopia's exports are almost entirely agricultural. Coffee alone is responsible for more than 50 percent of foreign-exchange earnings; other exported products are hides and skins, oilseeds, and vegetables. Major export destinations are the United States, western Europe, Japan, and Saudi Arabia. Manufactures, especially machinery and transport equipment, account for almost three-quarters of the value of imports; food products and fuels are also important. With more being spent on imports than is earned from exports, Ethiopia's balance of payments has been negative for many years.

Transport. Among the more successful developments in Ethiopia has been the road system. During the brief Italian occupation of 1935–41, highways were opened up linking Addis Ababa to the provinces, and after World War II the Imperial Highway Authority opened new feeder roads to isolated localities. Road construction and maintenance slowed during the decade of civil war in the 1980s. With the secession of Eritrea, Ethiopia no longer has direct access to the Red Sea ports of Aseb and Mitsiwa. This loss has placed greater importance on the railway between Addis Ababa and Djibouti, which was originally built by a French company between 1897 and 1917.

The air transport system

Ethiopia's air transport system has enjoyed a success unparalleled in Africa. The internal network of Ethiopian Airlines (a state-owned but independently operated carrier) is well developed, connecting major provincial capitals and locations of tourist interest. Its international network provides excellent service to Africa, Europe, and Asia. Bole International Airport, near Addis Ababa, serves other African and European airlines and is also an acknowledged centre for pilot training and aircraft maintenance.

Tourism. Although tourism was curtailed during the period of Derg rule, Ethiopia can once again realize the tourist potential of such historical wonders as the rock-hewn churches of Lalibela, the antiquities at Aksum, and the Gonder castles. Of equal attraction are Ethiopia's diverse peoples, their intriguing cultures, and the natural beauty of their land.

Administration and social conditions. *Government.* Ethiopia's ancient system of feudal government experienced significant changes under Haile Selassie I, who carefully grafted onto the traditional governing institutions a weak parliament of appointed and elected legislators, a judiciary with modernized civil and criminal codes and a hierarchy of courts, and an executive cabinet of ministers headed by a prime minister but answerable to himself. The Derg took power in 1974 and promised to bring revolutionary change to Ethiopia. Promulgating itself as the Provisional Military Administrative Council (PMAC) and later as the Workers' Party of Ethiopia (WPE), the Derg instituted a Soviet-style government with a state president and a house of deputies that were answerable to a revolutionary council with a politburo at the top. In May 1991 the Ethiopian People's Revolutionary Democratic Front (EPRDF) entered the capital. The EPRDF introduced a temporary constitution called the National Charter, created an 87-member assembly known as the State Council, and proceeded to form a cabinet for the Transitional Government of Ethiopia (TGE). The TGE endorsed the secession of Eritrea, realigned provincial boundaries in an attempt to create ethnic homogenates, demobilized the national armed forces, and suspended the courts and enforcing agencies.

Education. Ethiopia maintains two educational systems. The traditional system is rooted in Christianity and Islām. Christian education at the primary level is often conducted by clergy in the vicinity of places of worship. Higher education, with emphasis on traditional Christian dogma, is still run by most major centres of worship, the most prominent being monasteries in the northern and northwestern provinces. Graduation from these centres leads to a position within the priesthood and church hierarchy.

Modern education

Modern education was an innovation of the emperors Menilek II (reigned 1889–1913) and Haile Selassie I (1930–74), who established an excellent, but limited, system of primary and secondary education. In addition, colleges of liberal arts, technology, public health, building, law, social work, business, agriculture, and theology were opened in the 1950s and '60s. In 1961 Haile Selassie I University (now Addis Ababa University) was created to centralize the administration of higher education in the country. The Derg expanded primary education and gave university designation to the Agricultural College in Alemaya near Harer and to the Medical College in Gonder. Nevertheless, with only about 40 percent of the primary, 15 percent of the secondary, and 1 percent of the tertiary age groups enrolled, Ethiopia's educational system is still grossly inadequate.

Health and welfare. Ethiopia's health-care system includes primary health centres, clinics, and hospitals. Only major cities have hospitals with full-time physicians, and most of the hospitals are in Addis Ababa. With some 30,000 people per physician, access to modern health care is very limited. Fewer than 60 percent of births are attended by health staff, and infant mortality is approximately 130 per 1,000 live births. Life expectancy at birth is 47 years.

Most health facilities are government-owned. Hopes of increasing the number of Ethiopian doctors suffered during the Derg era, when many either left the country or failed to return from specialized training abroad. Two medical schools continue to produce general practitioners and a few specialists, but the scale of output does not match the rising demand. Under the Derg, health facilities deteriorated from lack of maintenance and from shortages of equipment and drugs. Widespread use of traditional healing continues to be important, including such specialized occupations as bonesetting, midwifery, and minor surgery (including circumcision).

Cultural life. The cultural heritage of Ethiopians resides in their religions, languages, and extended families. All major language and religious groups have their own cultural practices (which also vary by geographic location); however, there are commonalities that form strong and recognizable national traits. Most Ethiopians place less importance on artifacts of culture than they do on an idealized ethos of cultural refinement as reflected in a respect for human sanctity, the practice of social graces, and the blessings of accumulated wisdom. Religion provides the basic tenets of morality. The invocation of God is often all that is needed to seal agreements, deliver on promises, and seek justifiable redress. Hospitality is reckoned the ultimate expression of grace in social relations. Old age earns respect and prominence in society, especially because of the piety, wisdom, knowledge, prudence, and altruism that it is supposed to bestow.

Major holidays

The influence of the Ethiopian Orthodox church on the national culture has been strong. Easter (Amharic: Yetinsa-e Be-al, or Fassika), Christmas (Yelidet Be-al, or Genna), and the Finding of the True Cross (Meskel) have become dominant national holidays. In an effort to reduce the dominance of the Christian church, both the Derg and the current regime have elevated the status of Islām. Major Islāmic holidays include ʿĪd al-Fiṭr (ending the fast of Ramaḍān) and ʿĪd al-Aḍḥā (ending the period of pilgrimage to Mecca). (A.Me.)

For statistical data on the land and people of Ethiopia, see the *Britannica World Data* section in the BRITANNICA BOOK OF THE YEAR.

HISTORY

From prehistory to the Aksumite kingdom. That life is of great antiquity in Ethiopia is indicated by the Hadar remains, a group of skeletal fragments found in the lower Awash River valley. The bone fragments belong to *Australopithecus afarensis*, an apelike creature that lived about four million years ago and may have been an ancestor of modern humans.

Sometime between the 8th and 6th millennia BC, pastoralism and then agriculture developed in northern Africa and southwestern Asia, and, as the population grew, an ancient tongue spoken in this region fissured into the modern languages of the Afro-Asiatic, or Hamito-Semitic,

family. This family includes the Cushitic and Semitic languages now spoken in Ethiopia. During the 2nd millennium BC, cereal grains and the use of the plow were introduced into Ethiopia from the region of the Sudan, and a people speaking Ge'ez (a Semitic language) came to dominate the rich northern highlands of Tigray. There, in the 7th century BC, they established the kingdom of Da'amat. This kingdom dominated lands to the west, obtaining ivory, tortoiseshell, rhinoceros horn, gold, silver, and slaves and trading them to South Arabian merchants.

After 300 BC, Da'amat deteriorated as trade routes were diverted eastward for easier access to coastal ports. Subsequent wars of aggrandizement led to unification under the inland state of Aksum, which, from its base on the Tigray Plateau, controlled the ivory trade into the Sudan, other trade routes leading farther inland to the south, and the port of Adulis on the Gulf of Zula. Aksum's culture comprised Ge'ez, written in a modified South Arabian alphabet, sculpture and architecture based on South Arabian prototypes, and an amalgam of local and Middle Eastern dieties. Thus, evidence exists of a close cultural exchange between Aksum and the Arabian peninsula, but there are no scholarly grounds for the common belief that South Arabian immigrants actually peopled and created Aksum, even if many of them visited or even came to live there. Nevertheless, the ancient cultural exchange across the Red Sea became enshrined in Ethiopian legend in the persons of Makeda—the Queen of Sheba—and the Israelite king Solomon. Their mythical union was said to have produced Menilek I, the progenitor of Ethiopia's royal dynasty.

Contact with Arabia

By the 5th century Aksum was the dominant trading power in the Red Sea. Commerce rested on sound financial methods, attested to by the minting of coins bearing the effigies of Aksumite emperors. In the anonymous Greek travel book *Periplus Maris Erythraei*, written in the 1st century AD, Adulis is described as an "open harbour" containing a settlement of Greco-Roman merchants. It was through such communities, established for the purposes of trade, that the Monophysite Christianity of the eastern Mediterranean reached Ethiopia during the reign of Emperor Ezanas (*c.* 303–*c.* 350). By the mid-5th century, monks were evangelizing among the Cushitic-speaking Agew people to the east and south.

At its height, Aksum extended its influence northward to the southernmost reaches of Egypt, westward to the Cushite kingdom of Meroe, southward to the Omo River, and eastward to the spice coasts on the Gulf of Aden. Even the South Arabian kingdom of the Himyarites, across the Red Sea in what is now Yemen, came under the suzerainty of Aksum. In the early 6th century, Emperor Caleb (Ella-Asbeha; reigned *c.* 500–534) was strong enough to reach across the Red Sea in order to protect his coreligionists in Yemen against persecution by a Jewish prince. However, Christian power in South Arabia ended after 572, when the Persians invaded and disrupted trade. They were followed 30 years later by the Arabs, whose rise in the 7th and 8th centuries cut off Aksum's trade with the Mediterranean world.

The Zagwe and Solomonid dynasties. As Christian shipping disappeared from the Red Sea, Aksum's towns lost their vitality. The Aksumite state turned southward, conquering adjacent, grain-rich highlands. Monastic establishments moved even farther to the south—for example, a great monastery was founded near Lake Hayk in the 9th century. Over time, one of the subject peoples, the Agew, learned Ge'ez, became Christian, and assimilated their Aksumite oppressors to the point that Agew princes were able to transfer the seat of the empire southward to their own region of Lasta. Thus the Zagwe dynasty appeared in Ethiopia. Later ecclesiastical texts accused this dynasty of not having been of pure "Solomonid" stock (*i.e.*, not descended from the union of Solomon and the Queen of Sheba), but it was in the religious plane that the Zagwe nonetheless distinguished themselves. At the Zagwe capital of Roha, Emperor Lalibela (reigned *c.* 1185–1225) directed the hewing of 11 churches out of living rock—a stupendous monument to Christianity, which he and the other Zagwes fostered along with the Ethiopianization of the countryside.

The rock churches of Lalibela

The church hierarchy, however, continued to view the Zagwes with distaste, favouring instead the Amhara princes of northern Shewa, who claimed legitimacy as the avatars of the Aksumite dynasty. When Shewa's king Yekuno Amlak rebelled in 1270, he was supported by an influential faction of monastic churchmen, who condoned his regicide of Emperor Yitbarek and legitimated his descent from Solomon. The genealogy of the new Solomonid dynasty was published in the early 14th century in the *Kebra Negast* ("Glory of the Kings"), a pastiche of legends that related the birth of Menilek I, associated Ethiopia with the Judeo-Christian tradition, and provided a basis for Ethiopian national unity through the Solomonid dynasty, Shewan culture, and the Amharic language. Well-armed ideologically, the Ethiopian state was prepared for a struggle impending in its eastern and southern provinces, where Christianity was being pushed back by the forces of Islām.

Islāmic missionary preaching had led to the conversion of many pagan people living on the peripheries of Ethiopian rule. In the late 13th century, various Muslim sultanates on Ethiopia's southern border fell under the hegemony of Ifat, located on the eastern Shewan Plateau and in the Awash valley. Early in his reign (1314–44), the Ethiopian emperor Amda Tseyon marched southward, where he established strategic garrisons and divided jurisdictions into *gults*, or fiefs, whose holders paid an annual tribute. His heavy taxation of exports, especially of gold, ivory, and slaves that were transshipped from Ifat to Arabia, resulted in several rebellions led by Muslim sultans. Amda Tseyon and his successors replied with brutal pacification campaigns that carried Solomonid power into the Awash valley and even as far as Seylac (Zeila) on the Gulf of Aden.

Aggrandizement into non-Christian areas eventually stimulated an internal reform and consolidation of the Christian state. As heads of the church, Solomonid monarchs actively participated in the development of religious culture and discipline by building and beautifying churches, repressing pagan practices, and promoting the composition of theological and doctrinal works. Such close connection between church and state inevitably brought conflict. Because of the role played by the monasteries in the accession of the Solomonid dynasty, many of them had been given perpetual title to considerable landed benefices. Such power allowed the monasteries at times to intervene in disputes over succession to the Solomonid throne and even openly to fight the reigning monarch. On the other hand, the monk Abba Ewostatewos (*c.* 1273–1352) preached isolation from corrupting state influences and a return to Biblical teachings—including observance of the Judaic Sabbath on Saturday in addition to the Sunday observance, an idea deeply held by the rural masses. The great emperor Zara Yakob (reigned 1434–68) conceded the latter point in 1450 at the Council of Mitmak, but he also initiated severe reforms in the church, eliminating abuses by strong measures and executing the leaders of heretical sects. Zara Yakob also conducted an unsuccessful military campaign to annihilate the Beta Israel, or Falasha, a group of Agew-speaking Jews who practiced a non-Talmudic form of Judaism.

Zara Yakob valued national unity above all and feared Muslim encirclement. In 1445 he dealt Ifat such a crushing military defeat that hegemony over the Muslim states passed to the sultans of Adal, in the vicinity of Harer. About 1520 the leadership of Adal was assumed by Aḥmad ibn Ibrāhīm al-Ghāzi, a Muslim reformer who became known as Sahib al-Fath ("the Conqueror") to the Muslims and Aḥmad Grāñ ("Ahmad the Left-Handed") to the Christians. Aḥmad drilled his men in modern Ottomon tactics and led them on a jihad, or holy war, against Ethiopia, quickly taking areas on the periphery of Solomonid rule. In 1528 Emperor Lebna Denegel was defeated at the battle of Shimbra Kure, and the Muslims pushed northward into the central highlands, destroying settlements, churches, and monasteries. In 1541 the Portuguese, whose interests in the Red Sea were imperiled by Muslim power, sent 400 musketeers to train the Ethiopian army in European tactics. Emperor Galawdewos (reigned 1540–59) opted for a hit-and-run strategy and on Feb. 21, 1543, caught Aḥmad in the open near Lake Tana and

The invasions of Aḥmad Grāñ

killed him in action. The Muslim army broke, leaving the field and north-central Ethiopia to the Christians.

The Age of the Princes. Meanwhile, population pressures had mounted among the Oromo, a pastoral people who inhabited the upper basin of the Genalē (Jubba) River in what is now southern Ethiopia and northern Kenya. Oromo society was based upon an "age-set" system known as *gada,* in which all males born into an eight-year generation moved together through all the stages of life. The warrior classes *(luba)* raided and rustled in order to prove themselves, and in the 16th century they began to undertake long-distance expeditions into the recently depopulated Ethiopian Plateau, stopping only when blocked by physical obstacles or by military might. By 1600 the Oromo had spread so widely in Ethiopia that Emperor Sarsa Dengel (reigned 1563–97) limited his government to what are now Eritrea, the northern regions of Tigray and Gonder, and parts of Gojam, Shewa, and Welo. These constituted an easily defensible, socially cohesive unit that included mostly Christian, Semitic-speaking agriculturalists. The church continued to rail against the Oromo threat and exhorted its flock to restore Ethiopia to its ancient domains, but the Monophysite faith soon found itself facing a different kind of threat from Roman Catholicism.

Following close upon the Portuguese musketeers were missionaries who, sent by the Jesuit founder St. Ignatius of Loyola, sought to convert Ethiopia to the Western church. The most successful of these was the Jesuit Pedro Páez; his personal authority and eminent qualities were such that Emperor Susenyos (reigned 1607–32) was persuaded to accept the doctrine of the dual nature of Christ and to notify the pope of his acceptance. This apostasy attracted the elites but repulsed the masses and the monks, and Susenyos was forced to abdicate in favour of his son Fasilides (reigned 1632–67).

The Gonder period

Fasilides established a new capital at Gonder, a trading centre north of Lake Tana that connected the interior to the coast. At its height about 1700, the city supported the arts and educational, religious, and social institutions as well as Beta Israel craftspeople, Muslim traders, and a large Oromo population of farmers, day labourers, and soldiers. In order to protect Orthodox Christendom from the pagan Oromo, who were moving into southern Tigray and southeastern Gonder province, the monarchy turned to a newly assimilated Oromo aristocracy. Eventually the emperors at Gonder became little more than local magnates protected by Oromo generals. Meanwhile, agricultural development in the Gibē River basin was leading to the formation of Oromo states just south of Shewa, the Gonga people were developing their own states in the Kefa highlands on the west bank of the Omo River, and a line that claimed Solomonid descent was returning northern Shewa to Amhara rule. By the reign of Emperor Tekle Haimanot I (1706–08), little was left of the central government. The Zamana Masafent ("Age of the Princes"), 150 years of feudal anarchy, had commenced.

For most Ethiopians, life during the Age of the Princes was difficult. Everyone had a niche in society, few moved from class to class, and practically nobody questioned the social order. As armies traversed the highlands, ruining the countryside and forcing farmers off the land and onto the field of battle, the self-sufficient rural economy of the north broke down. The balance of power shifted southward to untouched Shewa, which prospered in the growing trade of the Gibē states. Shewa's self-proclaimed king, Sahle Selassie (reigned 1813–47), and his successors expanded southward, so that by 1840 they controlled most of Shewa to the Awash River and enjoyed suzerainty as far south as Guragē.

To the north, Kassa Hailu was ending the Age of the Princes. After serving as a mercenary in Gojam, Kassa returned to his native Kwara in the western lowlands, where he prospered as a highwayman and built a good, small army. By 1847 he monopolized the lowlands' revenues from trade and smuggling, forcing Gonder's leading magnates to integrate him into the establishment. Finally, in April 1853 at Takusa, Kassa defeated Ras (Prince) Ali, the last of the Oromo lords. After consolidating his rule over Tigray to the north, Kassa was crowned Emperor Tewodros II on Feb. 9, 1855. Later that year he marched south and forced the submission of Shewa.

Tewodros II, Yohannes IV, and Menilek II. Although Tewodros' first years were marked by attempts at social justice, his effort to establish garrisons nationwide lost the allegiance of the already heavily taxed peasantry, and he alienated parish clergy by nationalizing "excess" church land. Such political ineptitude gave heart to the regional aristocrats, who returned to conspiracy. The emperor held Ethiopia together only through coercion, and in 1861, finally understanding his conundrum, he conceived a bold foreign policy to regain popular support. In 1862 Tewodros offered Britain's Queen Victoria an alliance to destroy Islām. The British ignored the scheme, and, when no response came, Tewodros imprisoned the British envoy and other Europeans. This incident led to an Anglo-Indian military expedition in 1868. Sir Robert Napier, the commander, paid money and weapons to Kassa, a *dejazmatch* (earl) of Tigray, in order to secure passage inland, and on April 10, on the plains below Āmba Maryam (or Mek'dela), British troops defeated a small imperial force. In order to avoid capture, Tewodros committed suicide.

The Tigrayan Kassa took the imperial crown as Yohannes IV on Jan. 21, 1872. After ejecting two Egyptian armies from the highlands of Eritrea in 1875–76, Yohannes moved south, forcing Shewa's king Menilek to submit and to renounce imperial ambitions. Yohannes thus became the first Ethiopian emperor in 300 years to wield authority from Tigray south to Guragē. He then sought to oust the Egyptians from coastal Eritrea, where they remained after the Mahdists had largely taken over the Sudan, but he was unable to prevent Italy from disembarking troops at Mitsiwa (now Massawa) in February 1885. In order to weaken the emperor, Rome tried to buy Menilek's cooperation with thousands of rifles; the Shewan king remained faithful to Yohannes but took the opportunity in January 1887 to incorporate Harer into his kingdom. Meanwhile, Yohannes repulsed Italian forays inland, and in 1889 he marched into the Sudan to avenge Mahdist attacks on Gonder. On March 9, 1889, with victory in his grasp, he was shot and killed at Metema.

The Treaty of Wichale

Menilek declared himself emperor of Ethiopia on March 25, and at Wichale (or Ucciali, as the Italians called it) in Tigray on May 2 he signed a treaty of amity and commerce granting Italy rule over Eritrea. The Italian version of Article XVII of the Treaty of Wichale made Rome the medium for Ethiopia's foreign relations, whereas the Amharic text was noncommital. Learning that Rome had used the mistranslation to claim a protectorate over all of Ethiopia, Menilek first sought a diplomatic solution; meanwhile, during 1891–93, he sent expeditions south and east to obtain gold, ivory, musk, coffee, hides, and slaves to trade for modern weapons and munitions. In December 1895, after two years of good harvests had filled Ethiopia's granaries, Menilek moved his army into Tigray.

Rome believed that as few as 35,000 soldiers could control Ethiopia, but it was proved wrong on March 1, 1896, at the Battle of Adowa (Adwa), where General Oreste Baratieri led 14,500 Italian troops on a poorly organized attack against Menilek's well-armed host of some 100,000 fighters. The Italian lines crumbled, and at noon retreat was sounded. The emperor retired into Ethiopia to await negotiations, and on Oct. 26, 1896, he signed the Treaty of Addis Ababa, which abrogated the Treaty of Wichale.

Menilek subsequently directed the Solomonid state into areas never before under its rule. Between 1896 and 1906 Ethiopia expanded to its present size, taking in the highlands, the key river systems, and a buffer of low-lying arid or tropical zones around the state's central core. Revenues from the periphery were used to modernize the new capital of Addis Ababa, to open schools and hospitals, and to build communication networks. Menilek contracted with a French company to construct a railway between Addis Ababa and Djibouti, thus spurring the exploitation of the country's produce by foreign merchants in cooperation with the ruling elites.

The reign of Haile Selassie I. As Menilek aged, he appointed a cabinet to act for his grandson and heir designate, Iyasu, a son of the Oromo ruler of Welo. Upon

the emperor's death in 1913, Iyasu took power in his own right. Seeking a society free of religious and ethnic divisions, he removed many of Menilek's governors and integrated Muslims into the administration, outraging Ethiopia's Christian ruling class. During World War I, Iyasu dallied with Islām and with the Central Powers in the hope of regaining Eritrea. After the Allies formally protested, Addis Ababa's aristocrats met, accused Iyasu of apostasy and subversion, and deposed him on Sept. 27, 1916.

Iyasu was replaced by Menilek's daughter, Zauditu. Since it was considered unseemly for a woman to serve in her own right, Ras Tafari, the son of Ras Makonnen and a cousin of Menilek, served as Zauditu's regent and heir apparent. The prince developed a modern bureaucracy by recruiting the newly educated for government service. He also engineered Ethiopia's entry into the League of Nations in 1923, reasoning that collective security would protect his backward country from aggression. To brighten Ethiopia's external image, he hired foreign advisers for key departments and set about abolishing slavery—a process in which he was helped by Ethiopia's transition to a market economy.

By 1928, when Zauditu named Tafari king, the economy was booming, thanks mainly to the export of coffee. In the countryside, local officials built roads and improved communications, facilitating the penetration of traders and entrepreneurs. Ethiopians remained in charge of the economy, since Tafari forced foreigners to take local partners and maintained tight control over concessions.

On April 1, 1930, Zauditu died and Tafari declared himself emperor. He was crowned Haile Selassie I ("Strength of the Trinity"; his baptismal name) on November 2. In July 1931 the emperor promulgated a constitution that enshrined as law his prerogative to delegate authority to an appointed and indirectly elected bicameral parliament, among other modern institutions. During 1931–34 Haile Selassie instituted projects for roads, schools, hospitals, communications, administration, and public services. The combined effect of these projects was to open the country to the world economy. By 1932 revenues were pouring into Addis Ababa from taxes applied to 25,000 tons of coffee exported each year.

The Italian occupation

Haile Selassie's success convinced Italy's ruler Benito Mussolini to undertake a preemptive strike before Ethiopia grew too strong to oppose Italian ambitions in the Horn of Africa. After an Ethiopian patrol clashed with an Italian garrison at the Welwel oasis in the Ogaden in November–December 1934, Rome began seriously preparing for war. Haile Selassie continued to trust in the collective security promised by the League of Nations. Only on Oct. 2, 1935, upon learning that Italian forces had crossed the frontier, did he order mobilization. During the subsequent seven-month Italo-Ethiopian War, the Italian command used air power and poison gas to separate, flank, and destroy Haile Selassie's poorly equipped armies. The emperor went into exile on May 2, 1936.

For five years (1936–41) Ethiopia was joined to Eritrea and Italian Somaliland to form Italian East Africa. During this period Italy carried out a program of public works, concentrating especially on highways and on agricultural and industrial development. Resistance to the occupation continued, however. The Italians dominated the cities, towns, and major caravan routes, while Ethiopian patriots harried the occupiers and sometimes tested the larger garrison towns. When Italy joined the European war in June 1940, the United Kingdom recognized Haile Selassie as a full ally, and the emperor was soon in Khartoum to help train a British-led Ethiopian army. This joint force entered Gojam on Jan. 20, 1941, and encountered an enemy quick to surrender. On May 5 the emperor triumphantly returned to Addis Ababa. Ignoring the British occupation authorities, he quickly organized his own government.

In February 1945 at a meeting with U.S. president Franklin D. Roosevelt, Haile Selassie submitted memoranda stressing the imperative for recovering Eritrea and thereby gaining free access to the sea. In 1948 and again in 1949, two commissions established by the wartime Allies and by the United Nations reported that Eritrea lacked national consciousness and an economy that could sustain independence. Washington, wishing to secure a communications base in Asmera and naval facilities in Mitsiwa—and also to counter possible subversion in the region—supported Eritrea's federation with Ethiopia. The union took place in September 1952.

During the 1950s Ethiopia's coffee sold well in world markets. Revenues were used to centralize the government, to improve communications, and to modernize urban centres. In November 1955 the emperor promulgated a revised constitution, which permitted parliament to authorize finances and taxes, to question ministers, and to disapprove imperial decrees. The constitution also introduced an elected lower house of parliament, a theoretically independent judiciary, separation of powers, a catalog of human rights, and a mandate for bureaucratic responsibility to the people. At the same time, the emperor retained his power of decree and his authority to appoint the government. Among his ministers, he subtly established competing power factions—a stratagem that had the ultimate effect of retarding governmental cooperation, economic development, and bureaucratic modernization.

Opposition to the monarchy

Some educated officials concluded that the country would move ahead only when the imperial regime was destroyed. In December 1960, while the emperor was abroad, members of the security and military forces attempted a coup d'état. The coup rapidly unraveled, but not before the nation's social and economic problems were described in radical terms. Even then, the emperor ignored the coup's significance; in February 1961 he began to name a new government that reflected the status quo ante by depending on the landowning military, aristocracy, and oligarchy for authority. Haile Selassie showed himself unable to implement significant land reform, and as a result progressives and students opposed the regime. The monarchy gradually lost its credibility, especially as it became embroiled in intractable conflicts in Eritrea and with Somalia.

Somalia's independence in 1960 stimulated Somali nationalists in Ethiopia's Ogaden to rebel in February 1963. When Mogadishu joined the fighting, the Ethiopian army and air force smashed its enemy. Mogadishu's consequent military alliance with the Soviet Union upset the regional balance of power, driving up Ethiopia's arms expenditures and necessitating more U.S. assistance. Meanwhile, an insurrection in Eritrea, which had begun in 1960 mainly among Muslim pastoralists in the western lowlands, came to attract highland Christians disaffected by the government's dissolution of the federation in 1962 and the imposition of Amharic in the schools. At the same time, an increasingly radical student movement in Addis Ababa identified Haile Selassie as an agent of U.S. imperialism and his landowning oligarchs as the enemy of the people. Under the motto of "land to the tiller," the students sought to limit property size and rights, and, by fostering the Leninist notion that nationalities had the right to secede, they gave strength and ideological justification to the Eritrean rebellion.

By the early 1970s one-third of Ethiopia's 45,000 soldiers were in Eritrea, and others were putting down tax rebellions in Balē, Sidamo, and Gojam. In January 1974 there began a series of mutinies led by junior officers and senior noncommissioned officers, who blamed the imperial elites for their impoverishment and for the country's economic and social ills. For the government, the situation was greatly worsened by drought and famine in the overpopulated and overfarmed north, the denial of which became an international scandal. In June representatives of the mutineers constituted themselves as the Coordinating Committee (Derg) of the Armed Forces, Police, and Territorial Army. Major Mengistu Haile Mariam, of Harer's Third Division, was elected chairman. The Derg proceeded to dismantle the monarchy's institutions and to arrest Haile Selassie's cronies, confidantes, and advisers. It then campaigned against the old and senile emperor, who was deposed on Sept. 12, 1974. In November, after a shootout among members of the Derg, as many as 60 leaders of the old regime were executed. The new government issued a Declaration of Socialism on Dec. 20, 1974.

Socialist Ethiopia. The Derg borrowed its ideology from competing Marxist parties, one of which, the Ethiopian People's Revolutionary Party (EPRP), believed so strongly in civilian rule that it undertook urban guerrilla war against the military rulers. During the ensuing anarchy, Mengistu seized complete power as head of state. He then undertook class warfare against the EPRP, as a result of which thousands of Ethiopia's best-educated and idealistic young people were killed or exiled. At that moment of weakness, in May and June 1977, Somalia's army advanced into the Ogaden. Moscow labeled Mogadishu the aggressor and diverted arms shipments to Ethiopia, where Soviet and allied troops trained and armed a People's Militia, provided fighting men, and reequipped the army. Unable to entice the United States into resupplying its troops, Somalia withdrew in early 1978. Mengistu quickly shifted troops to Eritrea, where, by year's end, the secessionists were pushed back into mountainous terrain around Nak'fa.

The nationalized economy

Mengistu sought to transform Ethiopia into a command state led by a disciplined and loyal party that would control virtually every organ of authority. To this end a land-reform proclamation of 1975 transferred ownership of all land to the state and provided allotments of no more than 25 acres (10 hectares) to individual peasants who farmed the land themselves. Extensive nationalization of industry, banking, insurance, large-scale trade, and extra dwellings completed the reforms and wiped out the economic base of the old ruling class. To implement the reforms, adjudicate disputes, and administer local affairs, peasants' associations were organized in the countryside and precinct organizations (*kebellay*) in the towns. In 1984 the Workers' Party of Ethiopia was formed, with Mengistu as secretary-general, and in 1987 a new parliament inaugurated the People's Democratic Republic of Ethiopia, with Mengistu as president.

Farmers failed to generate the high yields expected from the land reform, mainly because the equitable distribution of land among the members of peasants' associations led to smaller plots, overcultivation, land degradation, and declining harvests. In order to feed Ethiopia's cities and the army, the government tried to force the peasants' associations to deliver grain at below-market prices—a disastrous policy that led to a horrific famine in 1984. With one-sixth of Ethiopia's people at risk of starvation, Western nations made available enough surplus grain to end the crisis by mid-1985. Donors were not so forthcoming for a mammoth population resettlement program that proposed to move people from the drought-prone and crowded north to the west and south, where supposedly surplus lands were available. The Mengistu regime handled the shift callously and did not have the necessary resources to provide proper housing, tools, medical treatment, or food for the 600,000 refugees it moved. Resources were also lacking for a related villagization program, which had the putative aim of concentrating scattered populations into villages where they might receive modern services. As late as 1990 most villages lacked the promised amenities because of resource-draining civil strife in the north.

By 1985–86 the government was embattled throughout most of Eritrea and Tigray, but Mengistu simply stepped up recruitment and asked Moscow for more arms. In December 1987 the EPLF broke through the Ethiopian lines before Nak'fa and continued to wage successful war with weapons captured from demoralized governmment troops. In early 1988 the EPLF began to coordinate its attacks with the Tigray People's Liberation Front (TPLF), which had long been fighting for the autonomy of Tigray and for the Marxist purity of the Ethiopian revolution. The Soviets refused to ship more arms, and in February 1989 a series of defeats and a worsening lack of weaponry forced the government to evacuate Tigray. The TPLF then organized the largely Amhara Ethiopian People's Democratic Movement. Together, these two groups formed the Ethiopian People's Revolutionary Democratic Front (EPRDF), and their forces easily advanced into Gonder and Welo provinces. The following year the EPLF occupied Mitsiwa; this broke the Ethiopian stranglehold on supplies entering the country and demonstrated that the government no longer ruled in Tigray and Eritrea. Shortly thereafter, when the TPLF cut the Addis Ababa–Gonder road and put Gojam at risk, Mengistu announced the end of socialism.

The peasantry immediately abandoned the regime's villages for their old homesteads, dismantled cooperatives, and redistributed land and capital goods. They ejected or ignored party and government functionaries, in several cases killing recalcitrant administrators. The regime was thus weakened in the countryside—especially in southern Ethiopia, where the long-dormant Oromo Liberation Front became active. By May 1991, with EPRDF forces controlling Tigray, Welo, Gonder, Gojam, and about half of Shewa, it was obvious that the army did not have sufficient morale, manpower, weapons, munitions, and leadership to stop the rebels' advance on Addis Ababa. Mengistu fled, and on May 28 the EPRDF took power.

Collapse of the Derg

The new government, led by a Tigrayan, the EPRDF chairman Meles Zenawi, claimed that it would democratize Ethiopia through recognition of the country's ethnic heterogeneity. No longer would the Ethiopian union be maintained by force; rather, it would be a voluntary federation of its many peoples. To this end the EPRDF and other political groups agreed to the creation of a transitional government that would engineer a new constitution and elections; to a national charter that recognized an ethnic division of political power; and to the right of nationalities to secede from Ethiopia (thus paving the way for Eritrea's legal independence in 1993).

It soon became obvious that the new regime was weakening the central administration of Ethiopia in favour of strong, ethnically based regional governments whose ruling parties were intended to be affiliated politically and ideologically with the EPRDF. A new regional map reflecting the changes was issued in 1992. Some Ethiopians criticized the new ethnic units as similar to an antinational reorganization of Ethiopia drawn up by the Italians in 1936. The Amhara, identified by the EPRDF as colonizers, were particularly affronted by the apparent disunification of the country. The government fought back by denouncing Amhara leaders as antidemocratic chauvinists and by muzzling the press through application of a new law that theoretically guaranteed its freedom. In the provinces, the government did not bother to maintain even the guise of freedom: there, the suppression of anti-EPRDF forces, especially the Oromo Liberation Front, was so blatant as to be noticed by the members of an international team sent to observe regional elections in June 1992. (H.G.M.)

Throughout 1992–93, the transitional government worked with donor governments and the World Bank to forge a structural adjustment program that devalued the Ethiopian currency, sharply reduced government intervention in the economy, and made it easier for foreign companies to invest in Ethiopia and repatriate their profits. However, the government was slow to denationalize land and to return property confiscated by the Derg, and, despite several years of healthy economic growth in the 1990s, there were regional food shortages in 1994 and 1999. The government repeatedly called upon the international donor community for help, but, by failing to free the economy and thus failing to solve its chronic famine crisis, Ethiopia remained unable to finance its own development and to create the surplus needed to relieve its own population.

Ethiopia promulgated a new constitution in 1995 that enshrined the EPRDF concept of regional and ethnic autonomy; yet the central government was able to maintain relatively strong control in most parts of the country, despite unrest from ethnic Somalis and some Islāmic groups. Eritrea maintained warm relations with Ethiopia following its secession, but these deteriorated rapidly in 1998, and war between the two states erupted in May of that year. Despite international efforts at mediation, a truce was not signed until December 2000, and the economic ramifications of this bloody conflict threatened Ethiopia's modest social and economic gains of the previous decade.

(Ed.)

For later developments in the history of Ethiopia, see the BRITANNICA BOOK OF THE YEAR.

Somalia

The Somali Democratic Republic (Somali: Jamhuuriyadda Dimuqraadiga Soomaaliya), a state in the Horn of Africa, occupies an important geopolitical position between sub-Saharan Africa and the countries of Arabia and southwestern Asia. With an area of about 246,000 square miles (637,000 square kilometres), it is bounded on the north by the Gulf of Aden and on the east by the Indian Ocean; from its southern point, its western border is bounded by Kenya and Ethiopia and, to the northwest, by Djibouti. The capital is Mogadishu (in Somali, Muqdisho or Xamar; in the colonial Italian rendering, Mogadiscio).

Living in a country of geographic extremes, with a dry and hot climate and a landscape of thornbush savanna and semidesert, the inhabitants of Somalia have developed equally demanding economic survival strategies. The Somali are Muslim, and about half follow a mobile way of life, pursuing nomadic pastoralism or agropastoralism. As a result, the Somali are an egalitarian, freedom-loving people who are suspicious of governmental authority.

PHYSICAL AND HUMAN GEOGRAPHY

The Somali Peninsula consists mainly of a tableland of young limestone and sandstone formations. Apart from a mountainous coastal zone in the north and several pronounced river valleys, most of the country is extremely flat, with few natural barriers to restrict the mobility of the nomads and their livestock.

Flat topography

The land. *Relief.* In the extreme north, along the Gulf of Aden, is a narrow coastal plain called the Guban. This gives way inland to a maritime mountain range with a steep, north-facing scarp. Near Ceerigaabo (Erigavo), a mountain called Surud Cad (Surud Ad) reaches the highest elevation in the country, about 7,900 feet (2,408 metres). To the south are the broad plateaus of the Galgodon (or Ogo) Highlands and the Sool and Hawd regions.

In southern Somalia the crystalline bedrock outcrops to the south of Baydhabo (Baidoa) in the shape of granite inselbergs. These give way farther south to alluvial plains, which are separated from the coast by a belt of dunes stretching more than 600 miles (1,000 kilometres) from south of Kismaayo (Chisimaio) to north of Hobyo (Obbia).

Drainage. The flatness of the Somalian plateaus is interrupted in the northeast by deep wadis, the Dharoor and Nugaal (Nogal) valleys. In the southwest are the only permanent rivers in Somalia, the Jubba (Giuba) and Shabeelle (Shebele), originating in the Ethiopian highlands. The Jubba carries more water than the Shabeelle, which sometimes dries up in its lower course in years of sparse rainfall in the Ethiopian highlands.

Soil. The types of soil vary according to climate and parent rock. The arid regions of northeastern Somalia have mainly thin and infertile desert soils. The limestone plateaus of the interfluvial area have some areas of fertile, dark gray to brown, calcareous residual soils that provide good conditions for rain-fed agriculture. The most fertile soils (vertisols, or black cotton soils) are found on the alluvial plains of the Jubba and Shabeelle rivers and are mainly used for irrigation agriculture.

Climate. Somalia lies astride the equator and so belongs to the tropics. Unlike typical climates at this latitude, conditions in Somalia range from arid in the northeastern and central regions to semiarid in the northwest and south.

The principal seasons

The climatic year comprises four seasons. The *gu,* or main rainy season, lasts from April to June; the second rainy season, called the *dayr,* extends from October to December. Each is followed by a dry season: the main one (*jilaal*) from December to March and the second one (*xagaa*) from June to September. During the second dry season, showers fall in the coastal zone.

Long-term mean annual rainfall is less than 4 inches (100 millimetres) in the northeast and approximately 8 to 12 inches in the central plateaus. The southwest and northwest receive an average of 20 to 24 inches a year. At Berbera on the northern coast the afternoon high averages more than 100° F (38° C) from June through September. The average afternoon high at Mogadishu ranges from 83° F (28° C) in July to 90° F (32° C) in April.

Plant and animal life. In accordance with rainfall distribution, southern and northwestern Somalia have a relatively dense thornbush savanna, with various succulents and species of acacia. By contrast, the high plateaus of northern Somalia have wide, grassy plains, with mainly low formations of thorny shrubs and scattered grass tussocks in the remainder of the region. Northeastern Somalia and large parts of the northern coastal plain, on the other hand, are almost devoid of vegetation. Exceptions to this are the wadi areas and the moist zones of the northern coastal mountains, where the frankincense tree (*Boswellia*) grows. The myrrh tree (*Commiphora*) thrives in the border areas of southern and central Somalia.

Owing to inappropriate land use, the original vegetation cover, especially in northern Somalia, has been heavily degraded and in various places even entirely destroyed. This progressive destruction of plant life also has impaired animal habitats and reduced forage, affecting not only Somalia's greatest resource, its livestock (chiefly goats, sheep, camels, and cattle), but also the wildlife. At present there are still many species of wild animals throughout the country—especially in the far south: hyenas, foxes, leopards, lions, warthogs, ostriches, small antelopes, and a large variety of birds.

Settlement patterns. Roughly half of the Somali population lives permanently in settled communities, the other half being nomadic pastoralists or agropastoralists. The settlements of the sedentary population consist of large, clustered villages near the rivers and in the central interfluvial area, as well as small hamlets farther away. The population is also concentrated in the old trading centres on the coast, such as Kismaayo, Baraawe (Brava), Marka (Merca), Mogadishu, Berbera, and Boosaaso (Bosaso).

The strong influence from Arabia, Persia, and India has shaped the face of the old coastal town centres, and Italian colonial architecture is visible in Mogadishu. There are two main types of traditional house: the typically African round house (*mundul*), mainly found in the interior, and the Arab-influenced rectangular house (*cariish*) with corrugated-steel roof, prevailing in the coastal regions and northern Somalia.

Pastoral nomads still live in transportable round huts called *aqal.* During the dry seasons, the high mobility of these livestock keepers leads to their temporary concentration in the river valleys of southern Somalia and around important water points all over the country.

Heavy migration from rural areas into towns has caused enormous urban expansion, especially in Mogadishu. As a result of increased market-oriented and extrapastoral activities, more nomads are tending to adopt a semi-settled

Michael S. Yamashita



Nomads drawing water from a well, Somalia.

way of life and economy. This has led to a great number of permanent nomad settlements, chiefly along the roads and tracks of the country's interior.

The people. In culture, language, and way of life, the people of Somalia, northeastern Kenya, the Ogaden region of Ethiopia, and the southern part of Djibouti are largely one homogeneous group.

Ethnic composition. The Somali people are divided into numerous clans, which are groups that trace their common ancestry back to a single father. These clans, which in turn are subdivided into numerous subclans, combine at a higher level to form clan families. The clan families inhabiting the interfluvial area of southern Somalia are the Rahanwayn and the Digil, which together are known as the Sab. Mainly farmers and agropastoralists, the Sab include both original inhabitants and numerous Somali groups that have immigrated into this climatically favourable area. Other clan families are the Daarood of northeastern Somalia, the Ogaden, and the border region between Somalia and Kenya; the Hawiye, chiefly inhabiting the area on both sides of the middle Shabeelle and south-central Somalia; and the Isaaq, who live in the central and western parts of northern Somalia. In addition, there are the Dir, living in the northwestern corner of the country but also dispersed throughout southern Somalia, and the Tunni, occupying the stretch of coast between Marka and Kismaayo. Toward the Kenyan border the narrow coastal strip and offshore islands are inhabited by the Bagiunis, a Swahili fishing people.

As well as the Somali, there is a sizable and economically important Bantu population, which is mainly responsible for the profitable irrigation agriculture practiced on the lower and middle reaches of the Jubba and Shabeelle rivers. Socially, however, they are regarded as inferior, many of them being descendants of former slaves. The result is a strict social distinction between the "noble" Somali of nomadic descent and the Bantu groups.

Arabs and Europeans

Another economically significant minority is the several tens of thousands of Arabs, mainly of Yemenite origin. By the end of the 1980s, the number of Italians permanently residing in Somalia (mainly as banana farmers) had dropped to only a few hundred.

Linguistic composition. The Somali language belongs to the Cushitic language family. Despite several regional dialects, it is understood throughout the country. The second official language is Arabic, which is spoken chiefly in northern Somalia and in the coastal towns. Owing to Somalia's colonial past, many people have a good command of English and Italian, which also are used in colleges and in the university. Swahili also is spoken in the south.

In 1973 Somalia adopted an official orthography based on the Latin alphabet. Until then, Somali had been an unwritten language.

Religion. Most Somali belong to the Shāfi'ī rite of the Sunnite sect of Islām. Various Muslim orders (*tariqa*) are important, especially the Qādirīyah, the Aḥmadīyah, and the Ṣaliḥiyah.

Demographic trends. The population of Somalia has been increasing annually by more than 3 percent, despite very high infant mortality and an average life expectancy of less than 50 years.

A high migration rate into the towns, chiefly by young men, has led to a disproportionately large percentage of old people in most rural areas and to high unemployment in the towns. Also, after the Ogaden conflict of 1977–78, hundreds of thousands of Somali from Ethiopia fled to Somalia, and during the ensuing civil war more than one million Somali sought shelter in neighbouring countries.

The economy. Somalia's economy is based on agriculture; however, the main economic activity is not crop farming but livestock raising. Between 1969 and the early 1980s, the military government imposed a system of "Scientific Socialism," which featured the nationalization of banks, insurance firms, oil companies, and all large industrial firms, the setting up of state-owned enterprises, farms, and trading companies, and the organizing of state-controlled cooperatives. In the end, this experiment weakened the Somalian economy considerably, and since the collapse of the military regime the economy has suffered even more as a result of civil war. Generally speaking, the Somalian economy cannot survive without foreign aid.

Resources. Somalia has few mineral resources—only some deposits of tin, phosphate, gypsum, guano, coal, iron ore, and uranium—but both quantity and quality are too low for mining to be worthwhile. However, the deposits of the clay mineral sepiolite, or meerschaum, in south-central Somalia are among the largest known reserves in the world. Exploitable oil and natural gas have not yet been found. Sea salt is collected at several sites on the coast.

Importance of livestock raising

Agriculture. By far the most important sector of the economy is agriculture, with livestock raising surpassing crop growing fourfold in value and earning about 90 percent of Somalia's foreign exchange. Agriculture in Somalia can be divided into three subsectors. The first is nomadic pastoralism, which is practiced outside the cultivation areas. This sector, which raises goats, sheep, camels, and cattle, has become increasingly market-oriented. The second sector is the traditional, chiefly subsistence, agriculture practiced by small farmers. This traditional sector takes two forms: rain-fed farming in the south and northwest, which raises sorghum, often with considerable livestock; and small irrigated farms along the rivers, which produce corn (maize), sesame, cowpeas, and, near towns, vegetables and fruits. The third sector consists of market-oriented farming on medium- and large-scale irrigated plantations along the lower Jubba and Shabeelle rivers. Here the major crops are bananas, sugarcane, rice, cotton, vegetables, grapefruit, mangoes, papayas, and other fruits.

Forestry. The acacia species of the thorny savanna in southern Somalia supply good timber and are the major source of charcoal, but charcoal production has long exceeded ecologically acceptable limits. More efficient and careful handling of *Boswellia, Commiphora,* and other resin-exuding trees could increase yields of aromatic gums.

Fishery. In the small fishing sector, tunny and mackerel are caught and canned in the north, sharks are often caught and sold dried by artisanal inshore fishers, and, in southern Somalia, choice fish and shellfish are processed for export.

Industry. In the late 1980s industry was responsible for just under 10 percent of Somalia's gross domestic product. Mogadishu was the chief industrial centre, with bottling plants, factories producing spaghetti, cigarettes, matches, and boats, a petroleum refinery, a small tractor assembly workshop, and small enterprises producing construction materials. In Kismaayo there were a meat-tinning factory, a tannery, and a modern fish factory. There were two sugar refineries, one near Jilib on the lower reach of the Jubba and one at Jawhar (Giohar) on the middle reach of the Shabeelle.

Low industrial production

Even before the destruction caused by Somalia's civil wars of the 1980s and '90s, the productivity of Somalian factories was very low. Often entire works did not operate at full capacity or produced nothing at all over long periods. A significant portion of commodities necessary for daily life is produced by small workshops in the informal sector.

Finance. The three principal banks, which are nationalized, are the Central Bank of Somalia, the Commercial and Savings Bank of Somalia, and the Somali Development Bank, which mainly provides loans for development projects. The currency, the Somali shilling, has been depreciating for years, and a shortage of hard currency greatly impedes the country's economic development.

Trade. Somalia has a large trade deficit. Its chief export commodities are livestock (to Arab countries, mainly Saudi Arabia) and bananas (to Italy and Arab countries). Other, much less important exports are hides and skins, fish, and frankincense and myrrh. Almost everything is imported, even food for an urban population no longer accustomed to the traditional diet.

Besides the official market, there is also a flourishing black market. Since wages in Somalia are very low, almost every family is directly or indirectly involved in informal trading.

Transportation. Inadequate transport facilities are a considerable impediment to Somalia's economic develop-

ment. There are no railways. Only about 1,800 miles of paved roads are passable year-round, and in the rainy seasons most rural settlements are not accessible by motor vehicle. Buses, trucks, and minibuses are the main means of transport for the population.

During peacetime, the state-owned Somali Airlines operated on national routes as well as on international routes to Kenya, Arabia, and Europe. Mogadishu, Berbera, and Kismaayo all have airports with long runways. These three cities also have deep-water harbours, but dangerous coral reefs keep coastal traffic to a minimum.

Administration and social conditions. *Government.* In 1960 Somalia became independent as a Western-style parliamentary democracy. A military coup in 1969, led by Major General Maxamed Siyaad Barre, inaugurated a phase of "Scientific Socialism." There existed one legal political party, the Somali Revolutionary Socialist Party, and various socialist-style mass organizations. Under the 1979 constitution the president and his supporters held the important positions of power, and a People's Assembly had no real power.

After years of destructive civil war waged by clan-based guerrillas, Siyaad's government fell in January 1991. In the north a de facto government declared the formation of an independent Somaliland Republic, while the fragmented, conflict-riven south lay largely in the hands of various clan militias. In 1993 the United Nations Operation in Somalia (UNOSOM II), consisting of military units from 26 countries and commanded by a special envoy from the United Nations, began an effort to stop the fighting and secure agreement on an interim government.

Education. Traditionally, Qur'ânic schools are responsible for the religious education of children according to Islâmic law. In addition, there is the state educational system, which, on the whole, is successful despite considerable shortcomings.

After primary school, students have a chance to attend agricultural secondary schools, a polytechnical school, a vocational training centre, a teachers' training centre, an agricultural college, or the National University, most of these institutions being located in and around Mogadishu.

Health and welfare. In spite of various health campaigns, the health service in Somalia is totally inadequate. Hospitals and dispensaries are located mainly in Mogadishu and, to a limited extent, in the district and provincial capitals.

Social services supplied by the state have fallen far short of demand. However, even in the towns the social network of family relationships is still intact and is of great help in times of need.

Cultural life. The varied cultural life of the Somali includes both traditional activities and, especially in the towns, many modern interests.

Daily life. Cultural activities primarily consist of poetry, folk dancing, the performance of plays, and singing. These traditional activities still retain their importance, especially in rural areas, and are practiced not only at family and religious celebrations but also at state ceremonies. On such occasions traditional local costume is generally worn.

Traditional and modern culture

Especially in the towns, traditional culture is rapidly being superseded by imported modern influences, such as television and videotapes, cinema, and bars and restaurants. Urban Somalian cooking has been strongly influenced by Italian cuisine, and young townspeople are much influenced by Western fashion in the way they dress. Association football (soccer) is a very popular sport.

The arts. There are many famous Somali artists, poets, musicians, actors, and dancers, some of whom live in exile. Nuruddin Farah, whose novels are written in English, has achieved international fame.

Cultural institutions in Mogadishu are the National Museum, the new Historical Museum, and the National Theatre. The Somali Academy of Sciences and Arts promotes research on Somalia.

Press and broadcasting. Press, radio, and television are all controlled and censored by the state. Books in general are hard to obtain, and the printing quality of the few books available in Somali is very poor. (J.H.A.J.)

For statistical data on the land and people of Somalia, see the *Britannica World Data* section in the BRITANNICA BOOK OF THE YEAR.

HISTORY

Before partition. *Early peoples of the coasts and hinterland.* From their connection with the Ethiopian hinterland, their proximity to Arabia, and their export of precious gums, ostrich feathers, ghee (clarified butter), and other animal produce as well as slaves from farther inland, the northern and eastern Somali coasts have for centuries been open to the outside world. This area probably formed part of Punt, "the land of aromatics and incense," mentioned in ancient Egyptian writings. Between the 7th and 10th centuries, immigrant Muslim Arabs and Persians developed a series of trading posts along the Gulf of Aden and Indian Ocean coasts. Islâm became firmly established in the northern ports of Seylac (Zeila) and Berbera and at Marka, Baraawe, and Mogadishu on the Indian Ocean coast in the south. These centres were engaged in a lively trade, with connections as far afield as China.

Probably by the 10th century the country from the Gulf of Aden coast inland was occupied first by Somali nomads and then, to their south and west, by various groups of pastoral Oromo who apparently had expanded from their traditional homelands in southwestern Ethiopia. To the south of these Cushitic-speaking Somali and Oromo—the "Berberi" of classical times and of the Arab geographers—the fertile lands between the Shabeelle and Jubba rivers were occupied, partly at least, by sedentary Bantu tribes of the Nyika confederacy, whose ancient capital was Shungwaya. Remnants of these Zanj, as they were known to the Arab geographers, still survive in this region, but their strongest contemporary representatives are found among the coastal Bantu, of whom the Pokomo live along the Tana River in northern Kenya. Another smaller allied population consisted of the ancestors of the scattered bands of hunters of northern Kenya and southern Somalia known as Wa-Ribi, or Wa-Boni, a people whose appearance and mode of existence recall those of the San.

The great Somali migrations. With this distribution of peoples in the 10th century, the stage was set for the great movements of expansion of the Somali toward the south and of the Oromo to the south and west. The first known major impetus to Somali migration was that of Sheikh Ismail Jabarti, ancestor of the Daarood Somali, who apparently came from Arabia to settle in the northeastern corner of the Somali Peninsula in the 11th century. This was followed, perhaps two centuries later, by the settlement of Sheikh Isaq, founder of the Isaaq Somali. As the Daarood and Isaaq clans grew in numbers and territory in the northeast, they began to vie with their Oromo neighbours, thus creating a general thrust toward the southwest. By the 16th century the movements that followed seem to have established much of the present distribution of Somali clans in northern Somalia. Other Somali pressed farther south, and some, according to the Arab geographer Ibn Said, had already reached the region of Marka by as early as the 13th century.

Somali pressure on the Oromo

In the meantime, farther to the west, a ring of militant Muslim sultanates had grown up around the Christian kingdom of Ethiopia, and the two sides were engaged in a protracted struggle for supremacy. Somali clansmen regularly formed part of the Muslim armies: the name Somali first occurs in an Ethiopian song of victory early in the 15th century. In the 16th century the Muslim state of Adal, whose port was Seylac, assumed the lead in the holy wars against the Christian Amhara. The turning point in the struggle between Christians and Muslims was reached with the Ethiopian victory in 1542, with Portuguese support, over the remarkable Muslim leader Aḥmad ibn Ibrâhîm al-Ghâzi, known to the Ethiopians as Aḥmad Grâñ. With his Somali armies, Aḥmad had harried Ethiopia almost to the point of collapse. This victory, which saved Ethiopia, also closed the door to Somali expansion westward and increased the pressure of the Somali and Oromo thrust southward. With this stimulus the main mass of the Oromo swept into Ethiopia from the south and southwest and streamed in conquering hordes as far north as the ancient city of Harer, which was laid to waste in 1567.

This massive invasion left something of a political vacuum in the south of the Horn, which new Somali settlers were quick to fill. By the 17th century the influx of new migrants, competing and jostling with each other, had become considerable. The old Ajuran Somali sultanate, linked with the port of Mogadishu, was overthrown and Mogadishu itself invaded and split into two rival quarters. Some of the earlier Somali groups found refuge in northern Kenya. The continuing Somali thrust south—largely at the expense of Oromo and Zanj predecessors—was ultimately only effectively halted at the Tana River by the establishment of administrative posts about 1912.

Somali clans and foreign traders. Thus, by the latter part of the 19th century the coastal and hinterland traditions had merged, and the centre of pressure had swung from the coast to the interior. In the north the ancient ports of Berbera and Seylac, much reduced in prosperity and importance, were now controlled by Somali nomads, and the position with the old ports of Marka, Baraawe, and Mogadishu was very similar. These towns had all been penetrated by various Somali clans, and the dominant political influence became that exercised by the Geledi clan ruling the lower reaches of the Shabeelle. Commercial and political links that provided an opening for European infiltration had, however, also been forged between these two coasts and the outside world. Part of the northern Somali coast including Seylac was then nominally under Turkish suzerainty, the Turkish claim going back to the 16th century, when Turkish forces had aided Aḥmad Grāñ in his campaigns against Ethiopia. The southern coastal towns, on the other hand, acknowledged the overlordship of the sultan of Zanzibar, although the latter's authority was slight in comparison with that exercised locally by the Geledi Somali.

The imperial partition. *Competition among the European powers and Ethiopia.* About the middle of the 19th century the Somali Peninsula became a theatre of competition between Great Britain, Italy, and France. On the African continent itself Egypt also was involved, and later Ethiopia, expanding and consolidating its realm under the guiding genius of the emperors Tewodros II, Yohannes IV, and Menilek II. Britain's interest in the northern Somali coast followed the establishment in 1839 of the British coaling station at Aden on the short route to India. The Aden garrison relied upon the importation of meat from the adjacent Somali coast. France sought its own coaling station and obtained Obock on the Afar coast in 1862, later thrusting eastward and developing the Somali port of Djibouti. Farther north, Italy opened a station in 1869 at Aseb, which, with later acquisitions, became the colony of Eritrea.

Stimulated by these European maneuvers, Egypt revived Turkey's ancient claims to the Red Sea coast. In 1870 the Egyptian flag was raised at Bullaxaar (Bulhar) and Berbera. With the disorganization caused by the revolt in the Sudan, however, Egypt was obliged to curtail its colonial responsibilities, evacuating Harer and its Somali possessions in 1885. In these circumstances the British government reluctantly decided to fill the gap left by Egypt. Between 1884 and 1886, accordingly, treaties of protection were drawn up with the main northern Somali clans guaranteeing them their "independence." Somali territory was not fully ceded to Britain, but a British protectorate was proclaimed and vice-consuls appointed to maintain order and control trade at Seylac, Berbera, and Bullaxaar. The interior of the country was left undisturbed.

Meanwhile, France had been assiduously extending its colony from Obock, and a clash with Britain was only narrowly averted when an Anglo-French agreement on the boundaries of the two powers' Somali possessions was signed in 1888. In the same period, the Italians were also actively extending their Eritrean colony and encroaching upon Ethiopian territory. Not to be outdone, Menilek took the opportunity of seizing the Muslim city of Harer, left independent after the Egyptian withdrawal. In 1889 Ethiopia and Italy concluded the Treaty of Wichale, which in the Italian view established an Italian protectorate over Ethiopia. Arms and capital were poured into the country, and Menilek was able to apply these new resources to bring pressure to bear on the Somali clansmen around Harer. In 1889 Italy also acquired two protectorates in the northeastern corner of Somalia; and by the end of the year the southern part of the Somali coast leased by the British East Africa Company from the sultan of Zanzibar was sublet to an Italian company.

Italian rule in the south

Italy had thus acquired a Somali colony. From 1892 the lease was held directly from Zanzibar for an annual rent of 160,000 rupees, and, after the failure of two Italian companies by 1905, the Italian government assumed direct responsibility for its colony of Italian Somaliland. To the south of the Jubba River the British East Africa Company held Jubaland until 1895, when this became part of Britain's East Africa protectorate. Britain and Italy reached agreement in 1884 on the extent of their respective Somali territories, but the Battle of Adwa (1896), at which the infiltrating Italian armies were crushed by Ethiopian forces, radically changed the position. Ethiopia, then independent of Italy, was plainly master of the hinterland, and in 1896–97 Italy, France, and Britain all signed treaties with Emperor Menilek, curtailing their Somali possessions. Italy gave up the Somali Ogaden, and Britain excised much of the western Hawd from its protectorate. Although the land and the Somali clansmen (who were not consulted), so abandoned, were not recognized as belonging to Ethiopia, there was nothing then to stop their gradual acquisition by Ethiopia.

Revolt in British Somaliland. These arrangements had scarcely been completed when the British Somaliland protectorate administration found its modest rule threatened by a religious rebellion led by Maxamed Cabdulle Xasan (in Arabic, Muḥammad ibn 'Abd Allāh Ḥasan). This Somali sheikh (known to the British as the Mad Mullah) of the Ogaadeen clan, living with his mother's people in the east of the protectorate, was an adherent of the Ṣaliḥiyah religious order, whose reformist message he preached with messianic zeal.

Sayyid Maxamed

Maxamed assumed the title of sayyid (lord), and his followers were known as the dervishes. He displayed great skill in employing all the traditional tactics of Somali clan politics in building up his following, strengthening these with the call to national Muslim solidarity against the infidel colonizers. Arms and ammunition, denied to Somali in the past, became easily available through the ports of Djibouti and the northeastern coast, and the dervishes, although opposed by many Somali, who were branded as traitors to Islām, successfully weathered four major British, Italian, and Ethiopian campaigns between 1900 and 1904. The cumbersome British armies, hampered by their supply and water requirements, found the dervish guerrilla tactics hard to combat effectively. A new policy was subsequently adopted, however, and, with the aid of an increasingly effective camel constabulary (whose founder, Richard Corfield, was killed at the Battle of Dulmadoobe in 1913), the dervishes were kept at bay until 1920, when a combined air, sea, and land operation finally routed them. The formidable dervish stronghold at Taleex, or Taleh, was bombed, but the sayyid escaped, as so often before, only to die of influenza a few months later while desperately seeking to rally his scattered followers.

Italian Somaliland. In Italian Somaliland, where the Italians had been gradually extending their hold on the country, the sayyid's rebellion had caused less disruption, and the appointment in 1923 of the first fascist governor marked a new active phase in the life of the colony. Two years later Britain ceded Jubaland with the port of Kismaayo, and in 1926, after a bitter military campaign, the two northern Italian protectorates were firmly incorporated. Italian settlement was encouraged, and fruit plantations were developed along the Shabeelle and Jubba valleys. Although agreements of 1897 and 1908 had defined the border with Ethiopia, this had not been demarcated, except for a stretch of about 18 miles delimited in 1910, and remained in dispute, thus facilitating the gradual Italian infiltration into Ethiopia. In 1934 the celebrated Welwel incident occurred in the eastern part of the Ogaden claimed by both Italy and Ethiopia. The Italian conquest of Ethiopia that followed brought the Ethiopian and Italian Somali territories together within the frame-

work of Italy's short-lived East African empire. Italian Somaliland became the province of Somalia.

The Somali Republic. *Independence and union.* During World War II the British protectorate was evacuated (1940) but was recaptured with Italian Somalia in 1941, when Ethiopia also was liberated. With the exception of French Somaliland, all the Somali territories were then united under British military administration. In 1948 the protectorate reverted to the Colonial Office; the Ogaden and the Hawd were gradually surrendered to Ethiopia; and in 1950 the Italians returned to southern Somalia with 10 years to prepare the country for independence under a United Nations trusteeship.

The British protectorate became independent on June 26, 1960. On July 1, Italian Somalia followed suit, and the two territories joined as the Somali Republic. The politics of the new republic were conditioned by clan allegiances, but the first major problems arose from the last-minute marriage between the former Italian trust territory and the former British protectorate. Urgent improvements in communication between the two areas were necessary, as were readjustments in their legal and judicial systems. The first independent government was formed by a coalition of the southern-based Somali Youth League (SYL) and the northern-based Somali National League (SNL).

Pan-Somalism. While modest developments were pursued internally with the help of mainly Western aid, foreign policy was dominated by the Somali unification issue and by the campaign for self-determination of adjoining Somali communities in the Ogaden, French Somaliland, and northern Kenya. The Somalian government strongly supported the Kenyan Somali community's aim of self-determination (and union with Somalia); when this failed in the spring of 1963, after a commission of inquiry endorsed Somali aspirations, Somalia broke off diplomatic relations with Britain and a Somali guerrilla war broke out in northern Kenya, paralyzing the region until 1967. By the end of 1963 a Somali uprising in the Ogaden led to a brief confrontation between Ethiopian and Somalian forces. Since the United States and the West provided military support to Ethiopia and Kenya, Somalia turned to the Soviet Union for military aid. Nevertheless, the republic maintained a generally neutral, but pro-Western, stance, and, indeed, a new government formed in June 1967 under the premiership of Maxamed Xaaji Ibrahiim Cigaal (Muhammad Haji Ibrahim Egal) embarked on a policy of détente with Kenya and Ethiopia, muting the Pan-Somali campaign.

The era of "Scientific Socialism." In March 1969 more than 1,000 candidates representing 64 parties (mostly clan-based) contested the 123 seats in the National Assembly. After these chaotic elections, all the deputies (with one exception) joined the SYL, which became increasingly authoritarian. The assassination of President Cabdirashiid Cali Sherma'arke (Abdirashid Ali Shermarke) on Oct. 15, 1969, provoked a government crisis, of which the military took advantage to stage a coup d'état on October 21.

Maxamed Siyaad Barre

The overthrow of Cigaal brought to power as head of state and president of a new Supreme Revolutionary Council the commander of the army, Major General Maxamed Siyaad Barre (Muhammed Siad Barre). At first the new regime concentrated on consolidating its power internally. Siyaad quickly adopted "Scientific Socialism," which, he claimed, was fully compatible with his countrymen's traditional devotion to Islām. Relations with socialist countries (especially the Soviet Union and China) were so greatly strengthened at the expense of Western connections that, at the height of Soviet influence, slogans proclaimed a trinity of "Comrade Marx, Comrade Lenin, and Comrade Siyaad." Rural society was integrated into this totalitarian structure through regional committees on which clan elders (now renamed "peace-seekers") were placed under the authority of a chairman, who was invariably an official of the state apparatus. Clan loyalties were officially outlawed, and clan-inspired behaviour became a criminal offense. Of the government's many "crash programs" designed to transform society, the most successful were mass literacy campaigns in 1973 and 1974, which made Somali a written language (in Latin characters) for the first time.

After 1974 Siyaad turned his attention to external affairs. Somalia joined the Arab League, gaining much-needed petrodollar aid and access to political support from those Persian Gulf states to which Somali labour and livestock were exported at a growing rate. Following Haile Selassie's overthrow in September 1974, Ethiopia began to fall apart, and guerrilla fighters of the Western Somali Liberation Front (WSLF) in the Ogaden pressed Siyaad (whose mother was an Ogaadeen) for support. When in June 1977 France granted independence to Djibouti (under a Somali president), the WSLF, backed by Somalia, immediately launched a series of fierce attacks on Ethiopian garrisons. By September 1977 the war was at the gates of Harer. Then the Soviet Union turned to fill the superpower vacuum left in Ethiopia by the gradual withdrawal of the United States. In the spring of 1978, with the support of Soviet equipment and Cuban soldiers, Ethiopia reconquered the Ogaden, and hundreds of thousands of Somali refugees poured into Somalia.

This terrible reversal strained the stability of the regime as the country faced a surge of clan pressures. An abortive military coup in April 1978 paved the way for the formation of two opposition groups: the Somali Salvation Democratic Front (SSDF), drawing its main support from the Majeerteen clan of the Mudug region in central Somalia, and the Somali National Movement (SNM), based on the Isaaq clan of the northern regions. Formed in 1982, both organizations undertook guerrilla operations from bases in Ethiopia. These pressures, in addition to pressure from Somalia's Western backers, encouraged Siyaad to improve relations with Kenya and Ethiopia. But a peace accord (1988) signed with the Ethiopian leader, Mengistu Haile Mariam, obliging each side to cease supporting Somali antigovernment guerrillas, had the ironic effect of precipitating civil war in Somalia.

Clan war. Threatened with the closure of their bases in Ethiopia, the SNM attacked government forces in their home region, while Ogaadeen Somali, who had been progressively absorbed into the army and militia, felt betrayed by the peace agreement with Ethiopia and began to desert, attacking Siyaad's clansmen. Siyaad became preoccupied with daily survival and consolidated his hold on Mogadishu. Clan-based guerrilla opposition groups multiplied rapidly, following the example of the SSDF and SNM. In January 1991, forces of the Hawiye-based United Somali Congress (USC) led a popular uprising that overthrew Siyaad and drove him to seek asylum among his own clansmen. Outside Mogadishu, all the main clans with access to the vast stores of military equipment in the country set up their own spheres of influence. In May 1991 the SNM, having secured control of the former British Somaliland northern region, declared an independent "Somaliland Republic." Government in the south had largely disintegrated and existed only at the local level in the SSDF-controlled northeast region. In Mogadishu the precipitate appointment of a USC interim government triggered a bitter feud between rival Hawiye clan factions. The forces of the two rival warlords, General Maxamed Farax Caydiid (Muḥammad Farah Aydid) and Cali Mahdi Maxamed (Ali Mahdi Muḥammad), tore the capital apart and battled with Siyaad's regrouped clan militia, the Somali National Front, for control of the southern coast and hinterland. This brought war and devastation to the grain-producing region between the rivers, spreading famine throughout southern Somalia. Attempts to distribute relief food were undermined by systematic looting and rake-offs by militias. In December 1992 the United States led a multinational force of more than 35,000 troops, which imposed an uneasy peace on the principal warring clans and pushed supplies into the famine-stricken areas. The military operation provided support for a unique effort at peacemaking by the United Nations. In January and March 1993, representatives of 15 Somali factions signed peace and disarmament treaties in Addis Ababa, but by June the security situation had deteriorated. American and European forces, suffering an unacceptable number of casualties, were withdrawn by March 1994. The UN force was reduced to military units from mainly Third World countries, and the clan-based tensions that had

The United Somali Congress

precipitated the civil war remained unresolved. (I.M.L.)

For later developments in the history of Somalia, see the BRITANNICA BOOK OF THE YEAR.

BIBLIOGRAPHY

The land and economy of eastern Africa. *Africa South of the Sahara* (annual) includes updated essays on all aspects of the countries of eastern Africa. Comprehensive overviews of Kenya, Uganda, and Tanzania are given in W.T.W. MORGAN, *East Africa* (1973); and W.T.W. MORGAN (ed.), *East Africa: Its Peoples and Resources,* 2nd ed. (1972). Plant life of the region is covered in E.M. LIND and M.E.S. MORRISON, *East African Vegetation* (1974); and D.J. PRATT and M.D. GWYNN (eds.), *Rangeland Management and Ecology in East Africa* (1977). The lakes are described in L.C. BEADLE, *The Inland Waters of Tropical Africa,* 2nd ed. (1981); and the rift valleys and their setting are explained in B.H. BAKER, P.A. MOHR, and L.A.J. WILLIAMS, *Geology of the Eastern Rift System of Africa* (1972). The anomalous climate of this portion of Africa is outlined in ch. 7 of GLENN T. TREWARTHA, *The Earth's Problem Climates,* 2nd ed. (1981).

Economic development is best treated in country-by-country accounts or continental overviews, such as *Accelerated Development in Sub-Saharan Africa: An Agenda for Action* (1982), a World Bank study. (W.T.W.M.)

The people of eastern Africa. GEORGE MURDOCK, *Africa: Its Peoples and Their Culture History* (1959), covers all the African peoples. JOCELYN MURRAY (ed.), *Cultural Atlas of Africa* (1981), is also helpful. A large series of thin but densely factual volumes, "Ethnographic Survey of Africa" (1950–), covers the entire region in two sections: "East Central Africa" and "North Eastern Africa."

East Africa: JOHN D. KESBY, *The Cultural Regions of East Africa* (1977), groups the peoples of East Africa and neighbouring areas into cultural regions and attempts to identify the processes by which cultural differences have arisen. LUCY MAIR, *Primitive Government: A Study of Traditional Political Systems in Eastern Africa,* rev. ed. (1977), studies the social organization of the best-documented of the peoples for the period 1890–1960, concentrating on obligations and the settling of disputes.

There are many studies of individual peoples. L.S.B. LEAKEY, *The Southern Kikuyu Before 1903,* 3 vol. (1977), is a detailed example, covering the southern third of one of the most numerous peoples of the Eastern Rift region. Other studies include ANDREW FEDDERS, *Peoples and Cultures of Kenya* (1979); JOHN LAMPHEAR, *The Traditional History of the Jie of Uganda* (1976); and J.C.D. LAWRANCE, *The Iteso: Fifty Years of Change in a Nilo-Hamitic Tribe of Uganda* (1957). (J.D.K.)

The Horn of Africa: JAN BRØGGER, *Belief and Experience Among the Sidamo* (1986), analyzes destiny, illness, and the spirit cult in Sidamo religion in relation to their economy. ASMAROM LEGESSE, *Gada: Three Approaches to the Study of African Society* (1973), gives the most comprehensive account of the *gada* system. DONALD N. LEVINE, *Wax & Gold: Tradition and Innovation in Ethiopian Culture* (1965, reprinted 1986), offers a thorough and detailed study of the culture and ethos of the politically dominant Amhara; his *Greater Ethiopia: The Evolution of a Multiethnic Society* (1974), explores cultural parallels and connections among the many different ethnic groups of the Ethiopian "mosaic." J. SPENCER TRIMINGHAM, *Islam in Ethiopia* (1952, reissued 1965), is a comprehensive cultural history of the Horn of Africa.

I.M. LEWIS, *Blood and Bone: The Call of Kinship in Somali Society* (1994), studies Somali society and culture and the linkages between "traditional" and modern political organization. (I.M.L.)

The history of eastern Africa. J.D. FAGE and ROLAND OLIVER (eds.). *The Cambridge History of Africa,* 8 vol. (1975–86), contains useful chapters on eastern African history, with comprehensive bibliographies. The contributions in UNESCO INTERNATIONAL SCIENTIFIC COMMITTEE FOR THE DRAFTING OF A GENERAL HISTORY OF AFRICA, *General History of Africa* (1981–), are also informative. (Ed.)

East Africa: BETHWELL A. OGOT (ed.), *Zamani: A Survey of East African History,* new ed. (1974), is still the best single-volume survey. ROLAND A. OLIVER *et al.* (eds.), *History of East Africa,* 3 vol. (1963–76), constitutes the most ambitious account so far. P.L. SHINNIE (ed.), *The African Iron Age* (1971), contains authoritative articles on archaeology by H.N. CHITTICK, "The Coast of East Africa," ch. 5, and by J.E.G. SUTTON, "The Interior of East Africa," ch. 6. G.S.P. FREEMAN-GRENVILLE, *The Medieval History of the Coast of Tanganyika* (1962), although subject now to correction, is still valuable. C.S. NICHOLLS, *The Swahili Coast: Politics, Diplomacy, and Trade on the East African Littoral, 1798–1856* (1971), is a very full study. FREDERICK COOPER, *Plantation Slavery on the East Coast of Africa* (1977), provides an excellent socioeconomic study of Zanzibar and Kenya in the 19th century. R.M.A. VAN ZWANENBERG and ANNE KING, *An Economic History of Kenya and Uganda, 1800–1970* (1975), concentrates on the years after 1900. (D.A.Lo./Ed.)

The Horn of Africa: The only book to deal with the history of this region is JOHN MARKAKIS, *National and Class Conflict in the Horn of Africa* (1987). Since there is no established historiography, the history of the entire Horn must be constructed from works devoted to Somalia and Ethiopia; see those sections below. I.M. LEWIS (ed.), *Nationalism & Self Determination in the Horn of Africa* (1983), discusses the rival ethnic nationalisms of the Horn, including those at the centre and periphery of Ethiopia. Works on the history of the Somali-Ethiopian conflict include TOM J. FARER, *War Clouds on the Horn of Africa: The Widening Storm,* 2nd rev. ed. (1979); and ROBERT F. KAPLAN, *Political Conflict on the Horn of Africa* (1981), which highlights the Ogaden war of 1977–78. (H.G.M.)

The countries of East Africa. *Kenya:* An up-to-date source of general information, *Kenya: An Official Handbook* (1988), was published on the 25th anniversary of independence. GUY ARNOLD, *Modern Kenya* (1981), gives a general survey of the country and its politics. FRANCIS F. OJANY and REUBEN B. OGENDO, *Kenya: A Study in Physical and Human Geography* (1973), D.C. EDWARDS and A.V. BOGDAN, *Important Grassland Plants of Kenya* (1951); and RICHARD S. ODINGO, *The Kenya Highlands: Land Use and Agricultural Development* (1971), analyze geographic and agricultural features. Population studies include S.H. OMINDE, ROUSHDI A. HENIN, and DAVID F. SLY (eds.), *Population and Development in Kenya* (1984), a useful focus on development implications of population growth; and S.H. OMINDE (ed.), *Kenya's Population Growth and Development to the Year 2000* (1988), an in-depth study. (S.H.O.)

BETHWELL A. OGOT (ed.), *Kenya Before 1900* (1976), contains essays on aspects of the history of various African peoples from about AD 500. ROBERT L. TIGNOR, *The Colonial Transformation of Kenya* (1976), recounts the impact of colonial rule on African individuals and societies from 1900 to 1939. MARJORIE RUTH DILLEY, *British Policy in Kenya Colony,* 2nd ed. (1966), explains the development of a British administrative philosophy to the mid-1930s. GUY ARNOLD, *Kenyatta and the Politics of Kenya* (1974), analyzes the role of Kenyatta in the political development of Kenya from 1922. DAVID THROUP, *Economic & Social Origins of Mau Mau, 1945–53* (1987), is well documented and researched. NORMAN N. MILLER, *Kenya: The Quest for Prosperity* (1984), carefully studies the politics, economics, and diplomacy of Kenya since independence. DONALD ROTHCHILD, *Racial Bargaining in Independent Kenya: A Study of Minorities and Decolonization* (1973), analyzes the conflict between ethnicity and national identity after independence. (K.In.)

Tanzania: Spatial aspects of resources and development are found in the official *Atlas of Tanzania,* 2nd ed. (1976); and the more comprehensive *Tanzania in Maps,* ed. by LEONARD BERRY (1971). ISSA G. SHIVJI, *Law, State, and the Working Class in Tanzania, c. 1920–1964* (1986), traces the creation of a working class during the colonial period. DEBORAH FAHY BRYCESON, *Food Insecurity and the Social Division of Labour in Tanzania, 1919–85* (1990), a thematic history of Tanzania, traces the development of the market, state, and client networks in relation to the fluctuation of the country's food supply. JANNIK BOESEN *et al.* (eds.), *Tanzania: Crisis and Struggle for Survival* (1986), collects articles on the rural economy written by a group of Scandinavian authors. ANDREW COULSON, *Tanzania: A Political Economy* (1982), is an interpretive account. (D.F.Br.)

I.N. KIMAMBO and A.J. TEMU (eds.), *A History of Tanzania* (1969), contains essays on political history from earliest times to independence. JOHN ILIFFE, *A Modern History of Tanganyika* (1979), is a comprehensive, documented, and scholarly general history from 1800 to 1961, mainly political but also economic and social, and *Tanganyika Under German Rule, 1905–1912* (1969), studies in detail the early colonial history. JOHN CHARLES HATCH, *Tanzania* (1972), by a sympathetic socialist writer, profiles the emergent country before and after independence. ANDREW ROBERTS (ed.), *Tanzania Before 1900* (1968), collects essays on the history of several ethnic groups before the colonial period. RODGER YEAGER, *Tanzania: An African Experiment,* 2nd ed., rev. and updated (1989), outlines the problems of independent Tanzania. JOHN GRAY, *History of Zanzibar, from the Middle Ages to 1856* (1962, reprinted 1975), offers a detached, scholarly study by a former chief justice of Zanzibar. ANTHONY CLAYTON, *The Zanzibar Revolution and Its Aftermath* (1981), gives a preliminary but acute assessment of the causes and immediate effects of the revolution of 1964. (K.In.)

Uganda: NELSON KASFIR, *The Shrinking Political Arena: Participation and Ethnicity in African Politics, with a Case Study of Uganda* (1976), examines ideas on ethnicity and ethnic categorizations in Uganda, drawing on case studies of ethnic

collision in the first Obote and Amin regimes. ALI A. MAZRUI, *Soldiers and Kinsmen in Uganda: The Making of a Military Ethnocracy* (1975), provides a fascinating if eclectic survey of modern Uganda, investigating the interplay between cultural, economic, and military forces. MICHAEL TWADDLE (ed.), *Expulsion of a Minority: Essays on Ugandan Asians* (1975), examines several aspects of Amin's summary expulsions of Asians. PETER LADEFOGED, RUTH GLICK, and CLIVE CRIPER, *Language in Uganda* (1972), is an interesting examination of the language scene. DUDLEY SEERS *et al.*, *The Rehabilitation of the Economy of Uganda*, 2 vol. (1979); and MARK BAIRD, *Uganda: Country Economic Memorandum* (1982), a World Bank study, detail economic development in independent Uganda. (O.H.K.)

SAMWIRI RUBARAZA KARUGIRE, *A Political History of Uganda* (1980), offers a brief, perceptive political history, mainly from 1860 to 1971. JAN JELMERT JØRGENSEN, *Uganda: A Modern History* (1981), is a scholarly account of the influence of the economy on Uganda's history from 1881 to 1979. T.V. SATHYAMURTHY, *The Political Development of Uganda, 1900–1986* (1986), details the history of colonial government and independence. M.S.M. SEMAKULA KIWANUKA, *A History of Buganda: From the Foundation of the Kingdom to 1900* (1971), chronicles the development of an African nation in a scholarly and well-researched manner. HOLGER BERNT HANSEN, *Mission, Church, and State in a Colonial Setting: Uganda, 1890–1925* (1984), carefully documents the important role played by Christian missionaries in the early years of British administration. MAHMOOD MAMDANI, *Politics and Class Formation in Uganda* (1976), gives a Marxist view of Uganda's history since the late 19th century. HOLGER BERNT HANSEN and MICHAEL TWADDLE (eds.), *Uganda Now: Between Decay and Development* (1988), collects essays on all aspects of Uganda's development since independence. (K.In.)

The countries of the Horn of Africa. *Djibouti:* Because most scholarship has been published in French, English-language sources for the geography and history of Djibouti are few and scattered. Among the fairly accessible articles and monographs in English on politics and economics are SAID YUSUF ABDI, "Independence for the Afars and Issas: Complex Background, Uncertain Future," *Africa Today*, 24(1):61–67 (January/March 1977), a succinct discussion of regional and internal politics at the time of independence; PETER D. COATS, "Factors of Intermediacy in Nineteenth-Century Africa: The Case of the Issa of the Horn," in THOMAS LABAHN (ed.), *Proceedings of the Second International Congress of Somali Studies*, vol. 2 (1984), pp. 175–199, an excellent analysis of the impact of the Franco-Ethiopian railway on the traditional trading networks and economy of the Issa Somali; and NORMAN N. MILLER, "The Other Somalia," *Horn of Africa*, 5(3):3–19 (1982), focusing on unrecorded trade between Somalia and Djibouti.

VIRGINIA THOMPSON and RICHARD ADLOFF, *Djibouti and the Horn of Africa* (1968), is the standard English-language text on the history of Djibouti, although it is now dated and lacks detail and depth of analysis; it can be supplemented by ROBERT THOLOMIER (ROBERT SAINT VÉRAN), *Djibouti, Pawn of the Horn of Africa*, abridged ed. (1981; originally published in French, 1977), covering the French Territory of the Afars and Issas from 1967 to 1977. The following French sources are also recommended: CENTRE DES HAUTES ÉTUDES SUR L'AFRIQUE ET L'ASIE MODERNES, *France, Océan Indien, Mer Rouge: études* (1986), with an extensive background chapter on Djibouti; PHILIPPE OBERLÉ and PIERRE HUGOT, *Histoire de Djibouti: des origines à la République* (1985), the most comprehensive history of Djibouti published to date; and OLIVIER WEBER (ed.), *Corne de l'Afrique* (1987), with several articles devoted to history, culture, and contemporary Djibouti life and with discussions developing the regional context. (C.C.C.)

Eritrea: A concise survey of ethnic groups living in Eritrea at the close of World War II is found in S.F. NADEL, *Races and Tribes of Eritrea* (1944). TEKESTE NEGASH, *Italian Colonialism in Eritrea, 1882–1941* (1987), surveys Italian colonialism and its impact on the people of Eritrea. G.K.N. TREVASKIS, *Eritrea: A Colony in Transition, 1941–52* (1960, reprinted 1975), recounts the political struggle over the fate of Eritrea in this time period. An account that reflects the Ethiopian point of view is HAGGAI ERLICH, *The Struggle Over Eritrea, 1962–1978* (1983). JOHN MARKAKIS, *National and Class Conflict in the Horn of Africa* (1987), includes an account of the Eritrean nationalist movement and the war of independence. JORDAN GEBRE-MEDHIN, *Peasants and Nationalism in Eritrea* (1989), analyzes Eritrean history from a nationalist perspective. (Jo.Ms.)

Ethiopia: EDWARD ULLENDORFF, *The Ethiopians: An Introduction to Country and People*, 3rd ed. (1973, reprinted 1990), is a comprehensive study; although some of the quantitative information is dated, the book makes interesting reading. MESFIN WOLDE-MARIAM, *An Introductory Geography of Ethiopia* (1972), is an excellent introduction providing information on the physical attributes, economic activities, population characteristics, and history of the country, and his *Rural Vulnerability to Famine in Ethiopia: 1958–1977* (1986), offers a valuable assessment, attempting to identify the human and natural causes of food insecurity. Two works on Ethiopia's peoples are those by Levine cited in the section above on the peoples of the Horn of Africa. DANIEL TEFERRA, *Social History and Theoretical Analyses of the Economy of Ethiopia* (1990), combines discussions on the historical geography of Ethiopia's people and on the current challenges in economic development. EDMOND J. KELLER, *Revolutionary Ethiopia: From Empire to People's Republic* (1988), deals with the political transformation of Ethiopia from its monarchist order to a people's republic. MULATU WUBNEH and YOHANNIS ABATE, *Ethiopia: Transition and Development in the Horn of Africa* (1988), is an excellent survey from a variety of perspectives, with discussions of the geography and history and of the country's social, cultural, political, and economic patterns. (A.Me.)

HAROLD G. MARCUS, *A History of Ethiopia* (1994), is the only modern general history of Ethiopia from *Australopithecus afarensis* to the fall of the Derg in 1991. Particular periods or events are covered in TADDESSE TAMRAT, *Church and State in Ethiopia, 1270–1527* (1972), which remains the only scholarly account of the golden years of the Solomonid dynasty; MOHAMMED HASSEN, *The Oromo of Ethiopia: A History, 1570–1860* (1990), the first modern history of the Oromo; MORDECHAI ABIR, *Ethiopia: The Era of the Princes: The Challenge of Islam and the Re-unification of the Christian Empire, 1769–1855* (1968), a dated but still largely accurate synthesis of the Age of the Princes; BAHRU ZEWDE, *A History of Modern Ethiopia, 1855–1974* (1991), a scholarly and authentically Ethiopian view; SVEN RUBENSON, *The Survival of Ethiopian Independence* (1976), an account that details the internal reasons for Ethiopia's continued independence during the epoch of modern European imperialism; CHRISTOPHER CLAPHAM, *Haile-Selassie's Government* (1969), an analysis of Haile Selassie's highly developed monarchical and authoritarian state; and JOHN W. HARBESON, *The Ethiopian Transformation: The Quest for the Post-Imperial State* (1988), a detailed account of the Mengistu years (1974–91) that argues against an Ethiopian revolution but for the notion of transformation. (H.G.M.)

Somalia: THOMAS LABAHN (ed.), *Proceedings of the Second International Congress of Somali Studies*, 4 vol. (1984), is a good collection of articles on socioeconomic development, politics, national science, and the arts. *Economic Transformation in a Socialist Framework* (1977), a report compiled by the JASPA Employment Advisory Mission, analyzes the economic changes under the socialist Somali government in the 1970s. JÖRG JANZEN, "Economic Relations Between Somalia and Saudi Arabia: Livestock Exports, Labor Migration, and the Consequences for Somalia's Development," *Northeast African Studies*, 8(2–3):41–51 (1986), analyzes the close economic interrelationship showing Somalia's dependence upon the Saudis. Janzen's "The Somali Inshore Fishing Economy: Structure, Problems, Perspectives," in ANNARITA PUGLIELLI (ed.), *Proceedings of the Third International Congress of Somali Studies* (1988), pp. 551–561, evaluates an important but under-exploited natural resource. JAN M. HAAKONSEN, *Scientific Socialism and Self Reliance* (1984), provides a well-informed account of the fishing cooperatives established for drought-afflicted pastoral nomads. PETER CONZE and THOMAS LABAHN (eds.), *Somalia: Agriculture in the Winds of Change* (1986), contains articles on the modern changes in crop production, pastoralism, and the socioeconomic environment. M.P.O. BAUMANN, JÖRG JANZEN, and H.J. SCHWARTZ (eds.), *Pastoral Production in Central Somalia* (1993), an interdisciplinary selection of articles, gives a profound insight into the structure and problems of Somalia's pastoral production. GARTH MASSY, *Subsistence and Change: Lessons of Agropastoralism in Somalia* (1987), explores the interfluvial area. ABDI ISMAIL SAMATAR, *The State and Rural Transformation in Northern Somalia, 1884–1986* (1989), is an analysis of the changes in rural areas. (J.H.A.J)

LEE V. CASSANELLI, *The Shaping of Somali Society: Reconstructing the History of a Pastoral People, 1600–1900* (1986), focuses on the history and society of southern Somalia. DAVID D. LAITIN and SAID S. SAMATAR, *Somalia: Nation in Search of a State* (1987), contains a general account of Somalian history, especially since independence in 1960. I.M. LEWIS, *A Modern History of Somalia: Nation and State in the Horn of Africa*, rev., updated, and expanded ed. (1988), is a comprehensive treatment of the political history of affairs in all the Somali territories. SAID S. SAMATAR, *Oral Poetry and Somali Nationalism* (1982), explains the crucial role of poetry in Somali politics, especially the case of nationalist leader Sayyid Maxamed Cabdulle Xasan, and his *Somalia: A Nation in Turmoil* (1991), provides a valuable overview of the factors leading to the collapse of the socialist state into clan-based warfare. AHMED I. SAMATAR, *Socialist Somalia: Rhetoric and Reality* (1988), analyzes the Siyaad regime's socialist policy of self-reliance and its efforts to achieve development. (I.M.L.)

Eastern Orthodoxy

Eastern Orthodoxy, characterized by its continuity with the apostolic church, its liturgy, and its territorial churches, is one of the three major branches of Christianity.

For coverage of related topics in the *Macropædia* and *Micropædia*, see the *Propædia*, section 827, and the *Index*.

This article is divided into the following sections:

Nature and significance 838
The cultural context
The norm of church organization
History 839
The church of imperial Byzantium
Orthodoxy under the Ottomans (1453–1821)
The church of Russia (1448–1800)
The Orthodox churches in the 19th century
The Orthodox church since World War I
Doctrine 848
Councils and confessions
God and man
Christ
The Holy Spirit
The Holy Trinity
The transcendence of God
Modern theological developments
The structure of the church 850
The canons
The episcopate
Clergy and laity
Monasticism
Worship and sacraments 851
The role of the liturgy
The eucharistic liturgies
The liturgical cycles
The sacraments
Architecture and iconography
The church and the world 854
Missions: ancient and modern
Orthodoxy and other Christians
Church, state, and society
Bibliography 856

Nature and significance

Eastern Orthodoxy is the large body of Christians who follow the faith and practices that were defined by the first seven ecumenical councils. The word orthodox ("right believing") has traditionally been used, in the Greek-speaking Christian world, to designate communities, or individuals, who preserved the true faith (as defined by those councils), as opposed to those who were declared heretical. The official designation of the church in Eastern Orthodox liturgical or canonical texts is "the Orthodox Catholic Church." Because of the historical links of Eastern Orthodoxy with the Eastern Roman Empire and Byzantium (Constantinople), however, in English usage it is referred to as the "Eastern" or "Greek Orthodox" Church. These terms are sometimes misleading, especially when applied to Russian or Slavic churches and to the Orthodox communities in western Europe and America. It should also be noted that there are Monophysitic churches (holding that after Incarnation Jesus had only a divine, and not a human and divine, nature) that have adopted the term orthodox as part of their names.

THE CULTURAL CONTEXT

The Schism of 1054

The schism between the churches of the East and the West (1054) was the culmination of a gradual process of estrangement that began in the first centuries of the Christian Era and continued through the Middle Ages. Linguistic and cultural differences, as well as political events, contributed to the estrangement. From the 4th to the 11th century, Constantinople, the centre of Eastern Christianity, was also the capital of the Eastern Roman, or Byzantine, Empire, while Rome, after the barbarian invasions, fell under the influence of the Holy Roman Empire of the West, a political rival. In the West theology remained under the influence of St. Augustine of Hippo (354–430), while in the East doctrinal thought was shaped by the Greek Fathers. Theological differences could have been settled if the two areas had not simultaneously developed different concepts of church authority. The growth of Roman primacy, based on the concept of the apostolic origin of the Church of Rome, was incompatible with the Eastern idea that the importance of certain local churches—Rome, Alexandria, Antioch, and later, Constantinople—could be determined only by their numerical and political significance. For the East, the highest authority in settling doctrinal disputes was the ecumenical council.

At the time of the Schism of 1054 between Rome and Constantinople, the membership of the Eastern Orthodox Church was spread throughout the Middle East, the Balkans, and Russia, with its centre in Constantinople, which was also called "New Rome." The vicissitudes of history have greatly modified the internal structures of the Orthodox Church, but, even today, the bulk of its members live in the same geographic areas. Missionary expansion toward Asia and emigration toward the West, however, have helped to maintain the importance of Orthodoxy worldwide.

THE NORM OF CHURCH ORGANIZATION

Autocephalous and autonomous churches

The Orthodox Church is a fellowship of "autocephalous" churches (governed by their own head bishops), with the Ecumenical Patriarch of Constantinople holding titular or honorary primacy. The number of autocephalous churches has varied in history. Today there are many: the Church of Constantinople (Istanbul), the Church of Alexandria (Egypt), the Church of Antioch (with headquarters in Damascus, Syria), and the churches of Jerusalem, Russia, Ukraine, Georgia, Serbia, Romania, Bulgaria, Cyprus, Greece, Albania, Poland, the Czech and Slovak republics, and America.

There are also "autonomous" churches (retaining a token canonical dependence upon a mother see) in Crete, Finland, and Japan. The first nine autocephalous churches are headed by "patriarchs," the others by archbishops or metropolitans. These titles are strictly honorary.

The order of precedence in which the autocephalous churches are listed does not reflect their actual influence or numerical importance. The patriarchates of Constantinople, Alexandria, and Antioch, for example, present only shadows of their past glory. Yet there remains a consensus that Constantinople's primacy of honour, recognized by the ancient canons because it was the capital of the ancient empire, should remain as a symbol and tool of church unity and cooperation. The modern pan-Orthodox conferences were thus convoked by the ecumenical patriarch of Constantinople. Several of the autocephalous churches are de facto national churches, by far the largest being the Russian Church; however, it is not the criterion of nationality but rather the territorial principle that is the norm of organization in the Orthodox Church.

Since the Russian Revolution there has been much turmoil and administrative conflict within the Orthodox Church. In western Europe and in the Americas, in particular, overlapping jurisdictions have been set up and political passions have led to the formation of ecclesiastical organizations without clear canonical status. Though it has provoked controversy, the establishment (1970) of the new autocephalous Orthodox Church in America by the patriarch of Moscow has as its stated goal the resumption of normal territorial unity in the Western Hemisphere.

History

THE CHURCH OF IMPERIAL BYZANTIUM

Byzantine Christianity about AD 1000. At the beginning of the 2nd millennium of Christian history, the church of Constantinople, capital of the Eastern Roman (or Byzantine) Empire, was at the peak of its world influence and power. Neither Rome, which had become a provincial town and its church an instrument in the hands of political interests, nor Europe under the Carolingian and Ottonian dynasties could really compete with Byzantium as centres of Christian civilization. The Byzantine emperors of the Macedonian dynasty had extended the frontiers of the empire from Mesopotamia to Naples (in Italy) and from the Danube River (in central Europe) to Palestine. The church of Constantinople not only enjoyed a parallel expansion but also extended its missionary penetration, much beyond the political frontiers of the empire, to Russia and the Caucasus.

Expansion of Byzantine church and state

Relations between church and state. The ideology that had prevailed since Constantine (4th century) and Justinian I (6th century)—according to which there was to be only one universal Christian society, the *oikoumenē*, led jointly by the empire and the church—was still the ideology of the Byzantine emperors. The authority of the patriarch of Constantinople was motivated in a formal fashion by the fact that he was the bishop of the "New Rome," where the emperor and the senate also resided (canon 28 of the Council of Chalcedon, 451). He held the title of "ecumenical patriarch," which pointed to his political role in the empire. Technically, he occupied the second rank—after the bishop of Rome—in a hierarchy of five major primates, which included also the patriarchs of Alexandria, Antioch, and Jerusalem. In practice, however, the latter three were deprived of all authority by the Arab conquest of the Middle East in the 7th century, and only the emerging Slavic churches attempted to challenge, at times, the position of Constantinople as the unique centre of Eastern Christendom.

The relations between state and church in Byzantium are often described by the term caesaropapism, which implies that the emperor was acting as the head of the church. The official texts, however, describe the emperor and the patriarch as a dyarchy (government with dual authority) and compare their functions to that of the soul and the body in a single organism. In practice, the emperor had the upper hand over much of church administration, though strong patriarchs could occasionally play a decisive role in politics: Patriarch Nicholas Mystikus (patriarch 901–907, 912–925) and Polyeuctus (patriarch 956–970) excommunicated emperors for uncanonical acts. In the area of faith and doctrine, the emperors could never impose their will when it contradicted the conscience of the church: this fact, shown in particular during the numerous attempts at union with Rome during the late medieval period, proves that the notion of caesaropapism is not unreservedly applicable to Byzantium.

The Church of the Holy Wisdom, or Hagia Sophia, built by Justinian in the 6th century, was the centre of religious life in the Eastern Orthodox world. It was by far the largest and most splendid religious edifice in all of Christendom. According to *The Russian Primary Chronicle*, the envoys of the Kievan prince Vladimir, who visited it in 987, reported: "We knew not whether we were in heaven or on earth, for surely there is no such splendor or beauty anywhere upon earth." Hagia Sophia, or the "great church," as it was also called, provided the pattern of the liturgical office, which was adopted throughout the Orthodox world. This adoption was generally spontaneous, and it was based upon the moral and cultural prestige of the imperial capital: the Orthodox Church uses the 9th-century Byzantine Rite.

Significance of Hagia Sophia

Monastic and mission movements. Both in the capital and in other centres, the monastic movement continued to flourish as it was shaped during the early centuries of Christianity. The Constantinopolitan monastery of Studion was a community of over 1,000 monks, dedicated to liturgical prayer, obedience, and asceticism. They frequently



The extent of Eastern Orthodoxy in the Balkans, the Middle East, eastern Europe, northern Asia, and Alaska.

opposed both government and ecclesiastical officialdom, defending fundamental Christian principles against political compromises. The Studite Rule (guidelines of monastic life) was adopted by daughter monasteries, particularly the famous Monastery of the Caves (Pecherskaya Lavra) in Kiev (in Russia). In 963 Emperor Nicephorus II Phocas offered his protection to St. Athanasius the Athonite, whose laura (large monastery) is still the centre of the monastic republic of Mt. Athos (under the protection of Greece). The writings of St. Symeon the New Theologian (949–1022), abbot of the monastery of St. Mamas in Constantinople, are a most remarkable example of Eastern Christian mysticism, and they exercised a decisive influence on later developments of Orthodox spirituality.

Missionary expansion

Historically, the most significant event was the missionary expansion of Byzantine Christianity throughout eastern Europe. In the 9th century, Bulgaria had become an Orthodox nation and under Tsar Symeon (893–927) had established its own autocephalous (administratively independent) patriarchate in Preslav. Under Tsar Samuel (976–1014) another autocephalous Bulgarian centre appeared in Ohrid. Thus, a Slavic-speaking daughter church of Byzantium dominated the Balkan Peninsula. It lost its political and ecclesiastical independence after the conquests of the Byzantine emperor Basil II (976–1025), but the seed of a Slavic Orthodoxy had been solidly planted. In 988 the Kievan prince Vladimir embraced Byzantine Orthodoxy and married a sister of Emperor Basil. After that time, Russia became an ecclesiastical province of the church of Byzantium, headed by a Greek or, less frequently, a Russian metropolitan appointed from Constantinople. This statute of dependence was not challenged by the Russians until 1448. During the entire period, Russia adopted and developed the spiritual, artistic, and social heritage of Byzantine civilization, which was received through intermediary Bulgarian translators. (See also below under *The church and the world—Missions: ancient and modern*).

Relations with the West. Relations with the Latin West, meanwhile, were becoming more ambiguous. On the one hand, the Byzantines considered the entire Western world as a part of the Roman *oikoumenē* of which the Byzantine emperor was the head and in which the Roman bishop enjoyed honorary primacy. On the other hand, the Frankish and German emperors in Europe were challenging this nominal scheme, and the internal decadence of the Roman papacy was such that the powerful patriarch of Byzantium seldom took the trouble of entertaining any relations with it. From the time of Patriarch Photius (patriarch 858–867, 877–886), the Byzantines had formally condemned the *Filioque* clause, which stated that the Holy Spirit proceeded from the Father and from the Son, as an illegitimate and heretical addition to the Nicene Creed, but in 879–880 Photius and Pope John VIII had apparently settled the matter to Photius' satisfaction. In 1014, however, the *Filioque* was introduced in Rome, and communion was broken again.

The background of the Schism of 1054

The incident of 1054, wrongly considered as the date of the Schism (which had actually been developing over a period of time), was, in fact, an unsuccessful attempt at restoring relations, disintegrating as they were because of political competition in Italy between the Byzantines and the Germans and also because of disciplinary changes (enforced celibacy of the clergy, in particular) imposed by the reform movement that had been initiated by the monks of Cluny, France. Conciliatory efforts of Emperor Constantine Monomachus (reigned 1042–55) were powerless to overcome either the aggressive and uninformed attitudes of the Frankish clergy, who were now governing the Roman Church, or the intransigence of Byzantine patriarch Michael Cerularius (1043–58). When papal legates came to Constantinople in 1054, they found no common language with the patriarch. Both sides exchanged recriminations on points of doctrine and ritual and finally hurled anathemas of excommunication at each other, thus provoking what has been called the Schism.

Invasions from East and West. *The Crusades.* After the Battle of Manzikert (1071) in eastern Asia Minor, Byzantium lost most of Anatolia to the Turks and ceased to be a world power. Partly solicited by the Byzantines, the Western Crusades proved another disaster: they brought the establishment of Latin principalities on former imperial territories and the replacement of Eastern bishops by a Latin hierarchy. The culminating point was, of course, the sack of Constantinople itself in 1204, the enthronement of a Latin emperor on the Bosporus, and the installation of a Latin patriarch in Hagia Sophia. Meanwhile, the Balkan countries of Bulgaria and Serbia secured national emancipation with Western help, the Mongols sacked Kiev (1240), and Russia became a part of the Mongol Empire of Genghis Khan.

The Byzantine heritage survived this series of tragedies mainly because the Orthodox Church showed an astonishing internal strength and a remarkable administrative flexibility.

Until the Crusades, and in spite of such incidents as the exchanges of anathemas between Michael Cerularius and the papal legates in 1054, Byzantine Christians did not consider the break with the West as a final schism. The prevailing opinion was that the break of communion with the West was due to a temporary take-over of the venerable Roman see by misinformed and uneducated German "barbarians," and that eventually the former unity of the Christian world under the one legitimate emperor—that of Constantinople—and the five patriarchates would be restored. This utopian scheme came to an end when the Crusaders replaced the Greek patriarchs of Antioch and Jerusalem with Latin prelates, after they had captured these ancient cities (1098–99). Instead of reestablishing Christian unity in the common struggle against Islām, the Crusades demonstrated how far apart Latins and Greeks really were from each other. When finally, in 1204, after a shameless sacking of the city, the Venetian Thomas Morosini was installed as patriarch of Constantinople and confirmed as such by Pope Innocent III, the Greeks realized the full seriousness of papal claims over the universal church: theological polemics and national hatreds were combined to tear the two churches further apart.

The sack of Constantinople and its aftermath

After the capture of the city, the Orthodox patriarch John Camaterus fled to Bulgaria and died there in 1206. A successor, Michael Autorianus, was elected in Nicaea (1208), where he enjoyed the support of a restored Greek empire. Although he lived in exile, this patriarch was recognized as legitimate by the entire Orthodox world. He continued to administer the immense Russian metropolitanate. From him, and not from his Latin competitor, the Bulgarian Church received again its right for ecclesiastical independence with a restored patriarchate in Trnovo (1235). It was also with the Byzantine government at Nicaea that the Orthodox Serbs negotiated the establishment of their own national church; their spiritual leader, St. Sava, was installed as autocephalous archbishop of Serbia in 1219.

The Mongol invasion. The invasion of Russia by the Mongols had disastrous effects on the future of Russian civilization, but the church survived, both as the only unified social organization and as the main bearer of the Byzantine heritage. The "metropolitan of Kiev and all Russia," who was appointed from Nicaea or from Constantinople, was a major political power, respected by the Mongol Khans. Exempt from taxes paid by the local princes to the Mongols and reporting only to his superior (the ecumenical patriarch), the head of the Russian Church—though he had to abandon his cathedral see of Kiev that had been devastated by the Mongols—acquired an unprecedented moral prestige. He retained ecclesiastical control over immense territories from the Carpathian Mountains to the Volga River, over the newly created episcopal see of Sarai (near the Caspian Sea), which was the capital of the Mongols, as well as over the Western principalities of the former Kievan Empire—even after they succeeded in winning independence (*e.g.*, Galicia) or fell under the political control of Lithuania and Poland.

Attempts at ecclesiastical union and theological renaissance. In 1261 the Nicaean emperor Michael Palaeologus recaptured Constantinople from the Latins, and an Orthodox patriarch again occupied the see in Hagia Sophia. From 1261 to 1453 the Palaeologan dynasty presided over an empire that was embattled from every side, torn apart by civil wars, and gradually shrinking to the very limits

Extension of ecclesiastical jurisdiction

of the imperial city itself. The church, meanwhile, kept much of its former prestige, exercising jurisdiction over a much greater territory, which included Russia as well as the distant Caucasus, parts of the Balkans, and the vast regions occupied by the Turks. Several patriarchs of this late period—*e.g.*, Arsenius Autorianus (patriarch 1255–59, 1261–65), Athanasius I (patriarch 1289–93, 1303–10), John Calecas (patriarch 1334–47), and Philotheus Coccinus (patriarch 1353–54, 1364–76)—showed great independence from the imperial power, though remaining faithful to the ideal of the Byzantine *oikoumenē*.

Without the military backing of a strong empire, the patriarchate of Constantinople was, of course, unable to assert its jurisdiction over the churches of Bulgaria and Serbia, which had gained independence during the days of the Latin occupation. In 1346 the Serbian Church even proclaimed itself a patriarchate; a short-lived protest by Constantinople ended with recognition in 1375. In Russia, Byzantine ecclesiastical diplomacy was involved in a violent civil strife; a fierce competition arose between the grand princes of Moscow and Lithuania, who both aspired to become leaders of a Russian state liberated from the Mongol yoke. The "metropolitan of Kiev and all Russia" was by now residing in Moscow, and often, as in the case of the metropolitan Alexis (1354–78), played a directing role in the Muscovite government. The ecclesiastical support of Moscow by the church was decisive in the final victory of the Muscovites and had a pronounced impact on later Russian history. The dissatisfied western Russian principalities (which would later constitute the Ukraine) could only obtain—with the strong support of their Polish and Lithuanian overlords—the temporary appointment of separate metropolitans in Galicia and Belorussia. Eventually, late in the 14th century, the metropolitan residing in Moscow again centralized ecclesiastical power in Russia.

Relations with the Western Church. One of the major reasons behind this power struggle in the northern area of the Byzantine world was the problem of relations with the Western Church. To most Byzantine churchmen, the young Muscovite principality appeared to be a safer bulwark of Orthodoxy than the Western-oriented princes who had submitted to Catholic Poland and Lithuania. Also, an important political party in Byzantium itself favoured union with the West in the hope that a new Western Crusade might be made against the menacing Turks. The problem of ecclesiastical union was, in fact, the most burning issue during the entire Palaeologan period.

The problem of ecclesiastical union

Emperor Michael Palaeologus (1259–82) had to face the aggressive ambition of the Sicilian Norman king Charles of Anjou, who dreamed of restoring the Latin empire in Constantinople. To gain the valuable support of the papacy against Charles, Michael sent a Latin-inspired confession of faith to Pope Gregory X, and his delegates accepted union with Rome at the Council of Lyons (1274). This capitulation before the West, sponsored by the Emperor, won little support in the church. During his lifetime, Michael succeeded in imposing an Eastern Catholic patriarch, John Beccus, upon the Church of Constantinople, but upon Michael's death an Orthodox council condemned the union (1285).

Throughout the 14th century, numerous other attempts at negotiating union were initiated by the emperors of Byzantium. Formal meetings were held in 1333, 1339, 1347, and 1355. In 1369 Emperor John V Palaeologus was personally converted to the Roman faith in Rome. All these attempts were initiated by the government and not by the church, for an obvious political reason; *i.e.*, the hope for Western help against the Turks. But the attempts brought no results either on the ecclesiastical or on the political levels. The majority of Byzantine Orthodox churchmen were not opposed to the idea of union but considered that it could only be brought about through a formal ecumenical council at which East and West would meet on equal footing, as they had done in the early centuries of the church. The project of a council was promoted with particular consistency by John Cantacuzenus, who, after a brief reign as emperor (1347–54), became a monk but continued to exercise great influence on all ecclesiastical and political events. The idea of an ecumenical council was initially rejected by the popes, but it was revived in the 15th century with the temporary triumph of conciliarist ideas (which advocated more power to councils and less to popes) in the West at the councils of Constance and Basel. Challenged with the possibility that the Greeks would unite with the conciliarists and not with Rome, Pope Eugenius IV called an ecumenical council of union in Ferrara, which later moved to Florence.

The Council of Ferrara–Florence

The Council of Ferrara–Florence (1438–45) lasted for months and allowed for long theological debates. Emperor John VIII Palaeologus, Patriarch Joseph, and numerous bishops and theologians represented the Eastern Church. They finally accepted most Roman positions—the *Filioque* clause, purgatory (an intermediate stage for the soul's purification between death and heaven), and the Roman primacy. Political desperation and the fear of facing the Turks again, without Western support, was the decisive factor that caused them to place their signatures of approval on the Decree of Union (July 6, 1439). The metropolitan of Ephesus, Mark Eugenicus, alone refused to sign. Upon their return to Constantinople, most other delegates also renounced their acceptance of the council and no significant change occurred in the relations between the churches.

The official proclamation of the union in Hagia Sophia was postponed until December 12, 1452; however, on May 29, 1453, Constantinople fell to the Ottoman Turks. Sultan Mehmed II transformed Hagia Sophia into an Islāmic mosque, and the few partisans of the union fled to Italy.

Theological and monastic renaissance. Paradoxically, the pitiful history of Byzantium under the Palaeologan emperors coincided with an astonishing intellectual, spiritual, and artistic renaissance that influenced the entire Eastern Christian world. The renaissance was not without fierce controversy and polarization. In 1337 Barlaam the Calabrian, one of the representatives of Byzantine Humanism, attacked the spiritual practices of the Hesychast (from the Greek word *hēsychia*, meaning quiet) monks, who claimed that Christian asceticism and spirituality could lead to the vision of the "uncreated light" of God. Barlaam's position was upheld by several other theologians, including Akyndinus and Nicephorus Gregoras. After much debate, the

Pierre Vals—Paris Match



Orthodox monk meditating in a chapel on Mount Athos.

church gave its support to the main spokesman of the monks, Gregory Palamas (1296–1359), who showed himself as one of the foremost theologians of medieval Byzantium. The councils of 1341, 1347, and 1351 adopted the theology of Palamas, and, after 1347, the patriarchal throne was consistently occupied by his disciples. John VI Cantacuzenus, who, as emperor, presided over the council of 1351, gave his full support to the Hesychasts. His close friend, Nicholas Cabasilas, in his spiritual writings on the divine liturgy and the sacraments, defined the universal Christian significance of Palamite theology. The influence of the religious zealots, who triumphed in Constantinople, outlasted the empire itself and contributed to the perpetuation of Orthodox spirituality under the Turkish rule. It also spread to the Slavic countries, especially Bulgaria and Russia. The monastic revival in northern Russia during the last half of the 14th century, which was associated with the name of St. Sergius of Radonezh, as well as the contemporaneous revival of iconography (*e.g.,* the work of the great painter Andrey Rublyov), would have been unthinkable without constant contacts with Mt. Athos, the centre of Hesychasm, and with the spiritual and intellectual life of Byzantium.

Along with the Hesychast revival, a significant "opening to the West" was taking place among some Byzantine ecclesiastics. The brothers Prochorus and Demetrius Cydones, under the sponsorship of Cantacuzenus, for example, were systematically translating the works of Latin theologians into Greek. Thus, major writings of Augustine, Anselm of Canterbury, and Thomas Aquinas were made accessible to the East for the first time. Most of the Latin-minded Greek theologians eventually supported the union policy of the emperors, but there were some—like Gennadios II Scholarios, the first patriarch under the Turkish occupation—who reconciled their love for Western thought with total faithfulness to the Orthodox Church.

ORTHODOXY UNDER THE OTTOMANS (1453–1821)

The Christian ghetto. According to Muslim belief, Christians, as well as Jews, were considered as "people of the Book"; *i.e.,* their religion was seen as not entirely false, but incomplete. Accordingly, provided that Christians submitted to the dominion of the caliphate and the Muslim political administration and paid appropriate taxes, they deserved consideration and freedom of worship. Any Christian mission or proselytism among the Muslims, however, was considered a capital crime. In fact, Christians were formally reduced to a ghetto existence: they were the *Rūm millet,* or the "Roman nation" conquered by Islām but enjoying a certain internal autonomy.

Conditions of Christians under Muslim rule

In January 1454 the Sultan allowed the election of a new patriarch, who was to become *millet-bachi,* the head of the entire Christian *millet,* or in Greek the "ethnarch," with the right to administer, to tax, and to exercise justice over all the Christians of the Turkish empire. Thus, under the new system, the patriarch of Constantinople saw his formal rights and jurisdiction extended both geographically and substantially: on the one hand, through the privileges granted to him by the sultan, he could practically ignore his colleagues, the other Orthodox patriarchs, and, on the other hand, his power ceased to be purely canonical and spiritual but became political as well. To the enslaved Greeks, he appeared not only as the successor of the Byzantine patriarchs but also as the heir of the emperors. For the Ottomans, he was the official and strictly controlled administrator of the *Rūm millet.* In order to symbolize these new powers, the patriarch adopted an external attire reminiscent of that of the emperors: mitre in form of a crown, long hair, eagles as insignia of authority, and other imperial accoutrements.

Consequences of the *millet* system

The new system had many significant consequences. Most important, it permitted the church to survive as an institution; indeed, the prestige of the church was actually increased because, for Christians, the church was now the only source of education and it alone offered possibilities of social promotion. Moreover, through the legal restrictions placed on mission, the new arrangement created the practical identification of church membership with ethnic origin. And finally, since the entire Christian *millet* was ruled by the patriarch of Constantinople and his Greek staff, it guaranteed to the Phanariots, the Greek aristocracy of the Phanar (now called Fener, the area of Istanbul where the patriarchate was, and still is, located), a monopoly in episcopal elections. Thus, Greek bishops progressively came to occupy all the hierarchical positions. The ancient patriarchates of the Middle East were practically governed by the Phanar. The Serbian and Bulgarian churches came to the same fate: the last remnants of their autonomy were formally suppressed in 1766 and 1767, respectively, by the Phanariot patriarch Samuel Hantcherli. This Greek control, exercised through the support of the hated Turks, was resented more and more by the Balkan Slavs and Romanians as the Turkish regime became more despotic, taxes grew heavier, and modern nationalisms began to develop.

It is necessary, however, to credit the Phanariots with a quite genuine devotion to the cause of learning and education, which they alone were able to provide inside the oppressed Christian ghetto. The advantages they obtained from the Porte (the Turkish government) for building schools and for developing Greek letters in the Romanian principalities of Moldavia and Walachia that were entrusted to their rule came to play a substantial role in the rebirth of Greece.

Relations with the West. The Union of Florence became fully inoperative as soon as the Turks occupied Constantinople (1453). In 1484 a council of bishops condemned it officially. Neither the sultan nor the majority of the Orthodox Greeks were favourable to the continuation of political ties with Western Christendom. The Byzantine cultural revival of the Palaeologan period was the first to experience adverse effects from the occupation. Intellectual dialogue with the West became impossible. Through liturgical worship and the traditional spirituality of the monasteries, the Orthodox faith was preserved in the former Byzantine world. Some self-educated men developed a remarkable ability to develop the Orthodox tradition through writings and publications, but they were isolated exceptions. Probably the most remarkable among them was St. Nicodemus of the Holy Mountain, the Hagiorite (1748–1809), who edited the famous *Philocalia,* an anthology of spiritual writings, and also translated and adapted Western spiritual writings (*e.g.,* those of the Jesuit founder, Ignatius of Loyola) into modern Greek.

East–West contacts

The only way for Orthodox Greeks, Slavs, or Romanians to acquire an education higher than the elementary level was to go to the West. Several of them were able to do so, but, in the process, became detached from their own theological and spiritual tradition.

The West, in spite of much ignorance and prejudice, had a constant interest in the Eastern Church. At times there was a genuine and respectful curiosity; in other instances, political and proselytistic (conversion) concerns prevailed. Thus, in 1573–81, a lengthy correspondence was initiated by the Lutheran scholars from Tübingen (in Germany). However interesting as a historical event, this correspondence, which includes the *Answers* of Patriarch Jeremias II (patriarch 1572–95), shows how little mutual understanding was possible at that time between the Reformers and traditional Eastern Christianity.

Relations with the West, especially after the 17th century, were often vitiated in the East by the incredible corruption of the Turkish government, which constantly fostered diplomatic intrigues. An outstanding example of such manipulation was the *kharāj,* an important tax required by the Porte at each patriarchal election. Western diplomats were often ready to provide the amount needed in order to secure the election of candidates favourable to their causes. The French and Austrian ambassadors, for example, supported candidates who would favour the establishment of Roman Catholic influence in the Christian ghetto, while the British and Dutch envoys supported patriarchs who were open to Protestant ideas. Thus, a gifted and Western-educated patriarch, Cyril Lucaris, was elected and deposed five times between 1620 and 1638. His stormy reign was marked by the publication in Geneva of a *Confession of Faith* (1629), which was, to the great amazement of all contemporaries, purely Calvinistic (*i.e.,* it contained Re-

formed Protestant views). The episode ended in tragedy. Cyril was strangled by Turkish soldiers at the instigation of the pro-French and pro-Austrian party. Six successive Orthodox councils condemned the *Confession:* Constantinople, 1638; Kiev, 1640; Jassy, 1642; Constantinople, 1672; Jerusalem, 1672; and Constantinople, 1691. In order to refute its positions, the metropolitan of Kiev, Peter Mogila, published his own *Orthodox Confession of Faith* (1640), which was followed, in 1672, by the *Confession* of the patriarch of Jerusalem, Dosítheos Notaras. Both, especially Peter Mogila, were under strong Latin influence.

These episodes were followed, in the 18th century, by a strong anti-Western reaction. In 1755 the Synod of Constantinople decreed that all Westerners—Latin or Protestant—had invalid sacraments and were only to be admitted into the Orthodox Church through Baptism. This practice of the Greek Church fell into disuse only in the 20th century.

THE CHURCH OF RUSSIA (1448–1800)

The "third Rome." *Origin of the Muscovite patriarchate.* At the Council of Florence, the Greek "metropolitan of Kiev and all Russia," Isidore, was one of the major architects of the Union. Having signed the decree, he returned to Moscow in 1441 as a Roman cardinal but was rejected by both church and state, arrested, and then allowed to escape to Lithuania. In 1448, after much hesitation, the Russians received a new primate, Jonas, elected by their own bishops. Their church became autocephalous, administratively independent under a "metropolitan of all Russia," residing in Moscow. In territories controlled by Poland, Rome (in 1458) appointed another "metropolitan of Kiev and all Russia." The tendencies toward separation from Moscow that had existed in the Ukraine since the Mongol invasion and that were supported by the kings of Poland thus received official sanction. In 1470, however, this metropolitan broke the union with the Latins and reentered—nominally—the jurisdiction of Constantinople, by then under Turkish control.

Independence of the Moscow metropolitanate

After this, the fate of the two churches "of all Russia" became quite distinct. The metropolitanate of Kiev developed under the control of Roman Catholic Poland. Hard pressed by the Polish kings, the majority of its bishops, against the will of the majority of their flock, eventually accepted union with Rome at Brest-Litovsk (1596). In 1620, however, an Orthodox hierarchy was reestablished, and a Romanian nobleman, Peter Mogila, was elected metropolitan of Kiev (1632). He created the first Orthodox theological school of the modern period, the famous Academy of Kiev. Modelled after the Latin seminaries of Poland, with instruction given in Latin, this school served as the theological training centre for almost the entire Russian high clergy in the 17th and 18th centuries. In 1686 the Ukraine was finally reunited with Muscovy, and the metropolitanate of Kiev was attached to the patriarchate of Moscow, with approval given by Constantinople.

Muscovite Russia, meanwhile, had acquired the consciousness of being the last bulwark of true Orthodoxy. In 1472 Grand Prince Ivan III (reigned 1462–1505) married Sofia (Zoë), the niece of the last Byzantine emperor. The Muscovite sovereign began to use more and more of the Byzantine imperial ceremonial, and he assumed the double-headed eagle as his state emblem. In 1510 the monk Philotheus of Pskov addressed Vasily III as "tsar" (or emperor), saying: "Two Romes have fallen, but the third stands, and a fourth there will not be." The meaning of the sentence was that the first Rome was heretical, the second—Byzantium—was under Turkish control, and the third was Moscow. Ivan IV, the Terrible, was crowned emperor, according to the Byzantine ceremonial, by the metropolitan of Moscow, Makary, on January 16, 1547. In 1551 he solemnly presided in Moscow over a great council of Russian bishops, the Stoglav ("Council of 100 Chapters"), in which various issues of discipline and liturgy were settled and numerous Russian saints were canonized. These obvious efforts to live up to the title of the "third Rome" lacked one final sanction: the head of the Russian Church was lacking the title of "patriarch." The "tsars" of Bulgaria and Serbia did not hesitate in the past to bestow the title on their own primates, but the Russians wanted an unquestionable authentication and waited for proper opportunity. It occurred in 1589 when the patriarch of Constantinople, Jeremias II, was on a fund-raising tour of Russia. He could not resist the pressure of his hosts and established the metropolitan Job as "patriarch of Moscow and all Russia." Confirmed later by the other Eastern patriarchs, the new patriarchate obtained the fifth place in the honorific order of the Oriental sees, after the patriarchs of Constantinople, Alexandria, Antioch, and Jerusalem.

The "third Rome" ideology

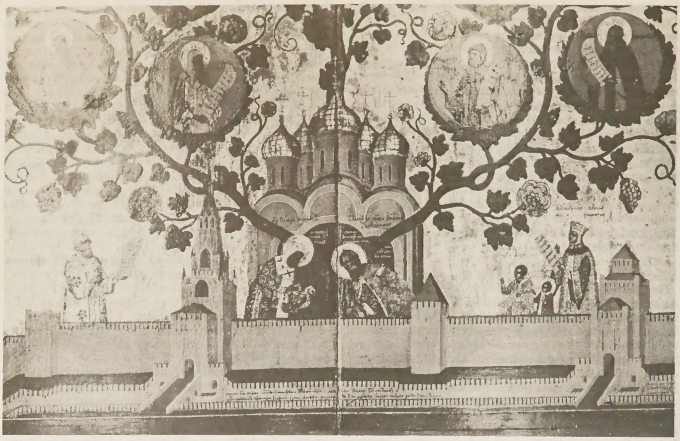
Relations between patriarch and tsar. After the 16th century, the Russian tsars always considered themselves as successors of the Byzantine emperors and the political protectors and financial supporters of Orthodoxy throughout the Balkans and the Middle East. The patriarch of Moscow, however, never pretended to occupy formally the first place among the patriarchs. Within the Muscovite Empire, many traditions of medieval Byzantium were faithfully kept. A flourishing monastic movement spread the practice of Christian asceticism in the northern forests, which were both colonized and Christianized by the monks. St. Sergius of Radonezh (*c.* 1314–92) was the spiritual father of this monastic revival. His contemporary, St. Stephen of Perm, missionary to the Zyryan tribes, continued the tradition of SS. Cyril and Methodius, the "apostles to the Slavs" in the 9th century, in translating the Scripture and the liturgy into the vernacular. He was followed by numerous other missionaries who promoted Orthodox Christianity throughout Asia and even established themselves on Kodiak Island off the coast of Alaska (1794). The development of church architecture, iconography, and literature also added to the prestige of the "third Rome."

Russian monastic and missionary activities

The Muscovite Empire, however, was quite different from Byzantium both in its political system and in its cultural self-understanding. The Byzantine "symphony" (harmonious relationship) between the emperor and the patriarch was never really applied in Russia. The secular goals of the Muscovite state and the will of the monarch always superseded canonical or religious considerations, which were still binding on the medieval emperors of Byzantium. Muscovite political ideology was always influenced more by the beginnings of western European secularism and by Asiatic despotism than by Roman or Byzantine law. Though strong patriarchs of Constantinople were generally able to oppose open violations of dogma and canon law by the emperors, their Russian successors were quite powerless; a single metropolitan of Moscow, St. Philip (metropolitan 1566–68), who dared to condemn the excesses of Ivan IV, was deposed and murdered.

A crisis of the "third Rome" ideology occurred in the middle of the 17th century. Nikon (reigned 1652–58), a strong patriarch, decided to restore the power and prestige of the church by declaring that the patriarchal office was superior to that of the tsar. He forced the tsar Alexis Romanov to repent for the crime of his predecessor against St. Philip and to swear obedience to the church. Simultaneously, Nikon attempted to settle a perennial issue of Russian church life: the problem of the liturgical books. Originally translated from the Greek, the books suffered many corruptions through the centuries and contained numerous mistakes. In addition, the different historical developments in Russia and in the Middle East had led to differences between the liturgical practices of the Russians and the Greeks. Nikon's solution was to order the exact compliance of all the Russian practices with the contemporary Greek equivalents. His liturgical reform led to a major schism in the church. The Russian masses had taken seriously the idea that Moscow was the last refuge of Orthodoxy. They wondered why Russia had to accept the practices of the Greeks, who had betrayed Orthodoxy in Florence and had been justly punished by God, in their view, by becoming captives of the infidel Turks. The reformist decrees of the patriarch were rejected by millions of lower clergy and laity who constituted the Raskol, or schism of the "Old Believers." Nikon was ultimately deposed for his opposition to the tsar, but his liturgical reforms were confirmed by a great council of the church

Schism in the Russian Church



"Planting the Tree of the Russian Nation," icon by Simon Ushakov, 1668. Detail shows the Kremlin, Ivan I Kalita, and Tsar Alexis. In the Tretyakov Art Gallery, Moscow.
Novosti Press Agency

that met in the presence of two Eastern patriarchs (1666–67).

The reforms of Peter the Great (reigned 1682–1725). The son of Tsar Alexis, Peter the Great, changed the historical fate of Russia by radically turning away from the Byzantine heritage and reforming the state according to the model of Protestant Europe. Humiliated by his father's temporary submission to Patriarch Nikon, Peter prevented new patriarchal elections after the death of Patriarch Adrian in 1700. After a long vacancy of the see, he abolished the patriarchate altogether (1721) and transformed the central administration of the church into a department of the state, which adopted the title of "Holy Governing Synod." An imperial high commissioner (*Oberprokuror*) was to be present at all meetings and, in fact, to act as the administrator of church affairs. Peter also issued a lengthy *Spiritual Regulation* (*Dukhovny Reglament*) that served as bylaws for all religious activities in Russia. Weakened by the schism of the "Old Believers," the church found no spokesman to defend its rights and passively accepted the reforms.

With the actions of Peter, the Church of Russia entered a new period of its history that lasted until 1917. The immediate consequences were not all negative. Peter's ecclesiastical advisers were Ukrainian prelates, graduates of the Kievan academy, who introduced in Russia a Western system of theological education; the most famous among them was Peter's friend, Feofan Prokopovich, archbishop of Pskov. Throughout the 18th century, the Russian Church also continued its missionary work in Asia and produced several spiritual writers and saints: St. Mitrofan of Voronezh (died 1703), St. Tikhon of Zadonsk (died 1783)—an admirer of the German Lutheran Johann Arndt and of German Pietism—as well as other eminent prelates and scholars such as Platon Levshin, metropolitan of Moscow (died 1803). All attempts at challenging the power of the tsar over the church, however, always met with failure. The metropolitan of Rostov, Arseny Matsiyevich, who opposed the secularization of church property by the empress Catherine the Great, was deposed and died in prison (1772). The atmosphere of secularistic officialdom that prevailed in Russia was not favourable for a revival of monasticism, but such a revival did take place through the efforts of a young Kievan scholar, Paissy Velichkovsky (1722–94), who became the abbot of the monastery of Neamts in Romania. His Slavonic edition of the *Philocalia* contributed to the revival of Hesychast traditions in Russia in the 19th century.

THE ORTHODOX CHURCHES IN THE 19TH CENTURY

Autocephalies in the Balkans. The ideas of the French Revolution, the nationalistic movements, and the everliving memory of past Christian empires led to the gradual disintegration of Turkish domination in the Balkans. According to a pattern existing since the late Middle Ages, the birth of national states was followed by the establishment of independent, autocephalous Orthodox churches. Thus the collapse of the Ottoman rule was accompanied by the rapid shrinking of the actual power exercised by the patriarch of Constantinople. Paradoxically, the Greeks, for whom—more than anyone—the patriarchate represented a hope for the future, were the first to organize an independent church in their new state.

The establishment of independent nationalistic churches

In Greece. In 1821 the Greek revolution against the Turks was officially proclaimed by the metropolitan of Old Patras, Germanos. The patriarchate, being the official Turkish-sponsored organ for the administration of the Christians, issued statements condemning and even anathematizing the revolutionaries. These statements, however, failed to convince anyone, least of all the Turkish government, which on Easter Day in 1821 had the ecumenical (Constantinopolitan) patriarch Gregory V hanged from the main gate of the patriarchal residence as a public example. Numerous other Greek clergy were executed in the provinces. After this tragedy, the official loyalty of the patriarchate was, of course, doubly secured. Unable either to communicate with the patriarchate or to recognize its excommunications, the bishops of liberated Greece gathered in Návplion and established themselves as the synod of an autocephalous church (1833). The ecclesiastical regime adopted in Greece was modelled after that of Russia: a collective state body, the Holy Synod, was to govern the church under strict government control. In 1850 the patriarchate was forced to recognize what was by then a fait accompli, and granted a charter of autocephaly (*tómos*) to the new Church of Greece.

In Serbia. The independence of Serbia led, in 1832, to the recognition of Serbian ecclesiastical autonomy. In 1879

the Serbian Church was recognized by Constantinople as autocephalous under the primacy of the metropolitan of Belgrade. This church, however, covered only the territory of what was called "old Serbia." The small state of Montenegro, always independent from the Turks, had its own metropolitan in Cetinje. This prelate, who was also the civil and military leader of the nation, was consecrated either in Austria, or, as in the case of the famous bishop-poet Pyotr II Negosh, in St. Petersburg (1833).

In the Austro-Hungarian empire, two autocephalous churches, with jurisdiction over Serbs, Romanians, and other Slavs, were in existence during the second half of the century. These were the patriarchate of Sremski-Karlovci (Karlowitz), established in 1848, which governed all the Orthodox in the Kingdom of Hungary; and the metropolitanate of Czernowitz (now Chernovtsy) in Bukovina, which, after 1873, also exercised jurisdiction over two Serbian dioceses (Zara and Kotor) in Dalmatia. The Serbian dioceses of Bosnia and Herzegovina, acquired by Austria in 1878, remained autonomous but were never completely independent from Constantinople.

In Romania. The creation of an independent Romania, after centuries of foreign control by Bulgarians, Turks, Greek-Phanariots, and, more recently, Russians, led in 1865 to the self-proclamation of the Romanian Church as autocephalous, even against the violent protests of the Phanar. As in Greece, the new church was under the strict control of the pro-Western government of Prince Alexandru Cuza. Finally, as in the Greek case, Constantinople recognized the Romanian autocephaly under the metropolitan of Bucharest (1885). The Romanians of Transylvania, still in Austria–Hungary, remained under the autocephalous metropolitan of Sibiu and others under the church of Czernowitz.

In Bulgaria. The reestablishment of the Church of Bulgaria eventually was secured, but not without tragedy and even a schism; this happened mainly because the issue of reestablishing the autocephalous church arose at a time when both Greek and Bulgarian populations lived side by side in Macedonia, Thrace, and Constantinople itself, though still within the framework of the Ottoman imperial system. After the Turkish conquest, and especially in the 17th and 18th centuries, the Bulgarians were governed by Greek bishops and were often prevented from worshipping in Slavonic. This enforced policy of Hellenization was rejected in the 19th century when Bulgarians began to claim not only a native clergy but also equal representation on the higher echelons of the Christian *millet*—*i.e.*, the offices of the patriarchate. These claims were met with firm resistance by the Greeks. The alternative was a national Bulgarian Church, which was created by a sultan's firman (decree) in 1870. The new church was to be governed by its own Bulgarian exarch, who resided in Constantinople itself and governed all the Bulgarians who recognized him. The new situation was uncanonical, because it sanctioned the existence of two separate ecclesiastical structures on the same territory. Ecumenical Patriarch Anthimus VI convened a synod in Constantinople, which also included the Greek patriarchs of Alexandria and Jerusalem (1872). The council condemned "phyletism"—the national or ethnic principle in church organization—and excommunicated the Bulgarians, who were certainly not alone guilty of "phyletism." This schism lasted until 1945, when a reconciliation took place with full recognition of Bulgarian autocephaly within the limits of the Bulgarian state.

The problem of the national or ethnic principle

After their liberation from the Turkish yoke, the Balkan churches freely developed both their national identities and their religious life. Theological faculties, generally following German models, were created in Athens, Belgrade (in Yugoslavia), Sofia (in Bulgaria), and Bucharest (in Romania). The Romanian Church introduced the full cycle of the liturgical offices in vernacular Romanian. But these positive developments were often marred by nationalistic rivalries. In condemning "phyletism," the synod of Constantinople (1872) had, in fact, defined a basic problem of modern Orthodoxy.

The church in imperial Russia. The *Spiritual Regulation* of Peter the Great remained in force until the very end of the Russian Empire (1917). Many Russian churchmen consistently complained against the submission of the church to the state, but there was little they could do except to lay plans for future reforms. This they did not fail to do, and in the 20th century the necessary changes were rapidly enacted. Though Peter himself and his first successors tended to deal personally and directly with church affairs, the tsars of the 19th century delegated much authority to the *Oberprokurors,* who received a cabinet rank in the government and were the real heads of the entire administration of the church. One of the most debilitating aspects of the regime was the legal division of Russian society by a rigid caste system. The clergy was one of the castes with its own school system, and there was little possibility for its children to choose another career.

In spite of these obvious defects, the church kept its self-awareness, and among the episcopate such eminent figures as Philaret of Moscow (1782–1867) promoted education, theological research, biblical translations, and missionary work. In each of its 67 dioceses, the Russian Church created a seminary for the training of priests and teachers. In addition, four theological academies, or graduate schools, were established in major cities (Moscow, 1769; St. Petersburg, 1809; Kiev, 1819; Kazan, 1842). They provided a generally excellent theological training for both Russians and foreigners. The rigid caste system and the strictly professional character of these schools, however, were obstacles to their seriously influencing society at large. It was, rather, through the monasteries and their spirituality that the church began to reach the intellectual class. More influential than the rigid discipline of the large monastic communities, the prophetic ministry of the "elders" (*startsy*), who acted as living examples of the standards of the spiritual life or as advisers and confessors, attracted large masses of the common people, and also intellectuals. St. Seraphim of Sarov (1759–1833), for example, lived according to the standards of the ancient Hesychast tradition that had been revived in the Russian forests. The *startsy* of Optino—Leonid (1768–1841), Makarius (1788–1860), and Ambrose (1812–91)—were visited not only by thousands of ordinary Christians but also by the writers Nikolay Gogol, Leo Tolstoy, and Fyodor Dostoyevsky. The latter was inspired by the *startsy* when he described in his novels monastic figures such as Zosima in *The Brothers Karamazov.* From the ranks of an emerging group of Orthodox lay intellectuals, the production of a living theology—if less scholarly than in the academies—was taking shape. The great influence of a lay theologian like Aleksey Khomyakov (1804–60), who belonged to the Slavophile (pro-Slavic) circle before it acquired a political flavour, eventually helped in the conversion to Orthodoxy at the end of the century of such leading Marxists as Sergey Bulgakov (1871–1944) and Nikolay Berdyayev (1874–1948). Missionary expansion also continued, particularly in western Asia, Japan, and Alaska (see below, *The church and the world: Missions: ancient and modern*).

Significance of Russian monasticism

Disproportionately larger and richer than its sister churches of the Balkans and the Middle East, the Church of Russia included, in 1914, more than 50,000 priests, 21,000 monks, and 73,000 nuns. It supported thousands of schools and missions. It cooperated with the Russian government in exercising great influence in Mid-Eastern affairs. Thus, with Russian help, an Arab (Meletios Doumani) rather than a Greek was elected for the first time as patriarch of Antioch (1899). With the successive partitions of Poland and the reunions with Russia of Belorussian and Ukrainian territories, many Eastern Catholic descendants of those who had joined the Roman communion in Brest-Litovsk (1596) returned to Orthodoxy.

The condition of the Russian Church on the eve of World War I

After 1905, Tsar Nicholas II gave his approval for the establishment of a preconciliar commission charged with the preparation of an all-Russian Church Council. The avowed goal of the planned assembly was to reestablish the church's independence, lost since Peter the Great, and eventually to restore the patriarchate. This assembly, however, was fated to meet only after the fall of the empire.

THE ORTHODOX CHURCH SINCE WORLD WAR I

The almost complete disappearance of Christianity in Asia Minor, the regrouping of the Orthodox churches in the

Balkans, the tragedy of the Russian Revolution, and the Orthodox diaspora in the West radically changed the entire structure of the Orthodox world.

The Russian Revolution and the Soviet period. The Church of Russia was less unprepared than generally believed to face the revolutionary turmoil. Projects of necessary reforms had been prepared since 1905, and most clergy did not feel particularly attached to the fallen regime that had deprived the church of its freedom for several centuries. During the rule of the provisional government, in August 1917 a council representing the entire church met in Moscow, including 265 members of the clergy and 299 laymen. The democratic composition and program of the council had been planned by the Pre-Conciliar Commission. It adopted a new constitution of the church that provided for the reestablishment of the patriarchate, the election of bishops by the dioceses, and the representation of laymen on all levels of church administration. It was only in the midst of the new revolutionary turmoil, however, that Tikhon, metropolitan of Moscow, was elected patriarch (October 31, 1917—six days after the Bolshevik takeover). The bloody events into which the country was plunged did not allow all the reforms to be carried out, but the people elected new bishops in several dioceses.

The effects of the Russian Revolution on the church

The Bolshevik government, because of its Marxist ideology, considered all religion as the "opium of the people." On January 20, 1918, it published a decree depriving the church of all legal rights, including that of owning property. The stipulations of the decree were difficult to enforce immediately, and the church remained a powerful social force for several years. The patriarch replied to the decree by excommunicating the "open or disguised enemies of Christ," without naming the government specifically. He also made pronouncements on political issues that he considered of moral importance: in March 1918 he condemned the peace of Brest-Litovsk that brought an unsatisfactory armistice between Russia and the Central Powers, and in October he addressed an "admonition" to Lenin, calling on him to proclaim an amnesty. Tikhon was careful, however, not to appear as a counterrevolutionary and in September 1919 called the faithful to refrain from supporting the Whites (anti-Communists) and to obey those decrees of the Soviet government that were not contrary to their Christian conscience.

The independence of the church suffered greatly after 1922. In February of that year, the government decreed the confiscation of all valuable objects preserved in the churches. The patriarch would have agreed to that measure if he had had the means to check on the government contention that all confiscated church property would be used to help the starving population on the Volga. The government refused all guarantees but supported a group of clergy who were ready to cooperate with it and to overthrow the patriarch. While Tikhon was under house arrest, this group took over his office and soon claimed the allegiance of a sizable proportion of bishops and clergy. This became known as the schism of the "Renovated" or "Living" Church, and it broke the internal unity and resistance of the church. Numerous bishops and clergy faithful to the patriarch were tried and executed, including the young and progressive metropolitan Benjamin of Petrograd. The "Renovated" Church soon broke the universal discipline of Orthodoxy by admitting married priests to the episcopate and by permitting widowed priests to remarry.

Upon his release, Tikhon condemned the schismatics, and many clergy returned to his obedience. But he also published a declaration affirming that he "was not the enemy of the Soviet government" and dropped any opposition to the authorities. Tikhon's attitude of conformism did not bring immediate results. His designated successors (after he died in 1925) were all arrested. In 1927 the "substitute *locum tenens*" (holder of the position) of the patriarchate, Metropolitan Sergius, pledged loyalty to the Soviet government. Nevertheless, under the rule of Joseph Stalin in the late 1920s and '30s, the church suffered a bloody persecution that claimed thousands of victims. By 1939 only three or four Orthodox bishops and 100 churches could officially function: the church was practically suppressed.

A spectacular reversal of Stalin's policies occurred, however, during World War II. Sergius was elected patriarch in 1943 and the "Renovated" schism was ended. Under Sergius' successor, Patriarch Alexis (1945–70), the church was able to open 25,000 churches and the number of priests reached 33,000. But a new antireligious move was initiated by Prime Minister Nikita Khrushchev in 1959–64, reducing the number of open churches to less than 10,000. Patriarch Pimen was elected in 1971 following Alexis' death, and, although the church still commanded the loyalty of millions, its future remained uncertain.

Post-World War II developments in the Soviet Union

After 70 years of repression and antireligious propaganda, however, the church experienced greater religious freedom in the late 1980s, culminating with the dissolution of the Soviet Union in 1991.

The Balkans and eastern Europe. In bringing about the fall of the Turkish, Austrian, and Russian empires, World War I provoked significant changes in the structures of the Orthodox Church. On the western borders of what was then the Soviet Union, in the newly born republics of Finland, Estonia, Latvia, and Lithuania, the Orthodox minorities established themselves as autonomous churches. The first three joined the jurisdiction of Constantinople, and the Lithuanian diocese remained nominally under Moscow. In Poland, which then included several million Belorussians and Ukrainians, the ecumenical patriarch established an autocephalous church (1924) over the protests of Patriarch Tikhon. After World War II the Estonian, Latvian, and Lithuanian autonomies were again suppressed, and in Poland the Orthodox Church was first reintegrated to the jurisdiction of Moscow and later was declared autocephalous again (1948).

In the Balkans, changes were even more significant. The five groups of Serbian dioceses (Montenegro, patriarchate of Karlovci, Dalmatia, Bosnia-Herzegovina, Old Serbia) were united (1920–22) under one Serbian patriarch, residing in Belgrade, the capital of the new Yugoslavia. Similarly, the Romanian dioceses of Moldavia-Walachia, Transylvania, Bukovina, and Bessarabia formed the new patriarchate of Romania (1925), the largest autocephalous church in the Balkans. Finally, in 1937, after some tension and a temporary schism, the patriarchate of Constantinople recognized the autocephaly of the Church of Albania.

Relations between church and state in the Balkans

After World War II, Communist regimes were established in the Balkan states. There were no attempts, however, at liquidating the churches entirely, similar to the persecutions that took place in Russia in the 1920s and '30s. In both Yugoslavia and Bulgaria, church and state were legally separated. In Romania, paradoxically, the Orthodox Church remained legally linked to the Communist state. With its solid record of resistance to the Germans, the Serbian Church was able to preserve more independence from the government than its sister churches of Bulgaria and Romania. Generally speaking, however, all the Balkan churches adopted an attitude of loyalty to the new regime, according to the pattern given by the patriarchate of Moscow. At that price, they could keep some theological schools, some publications, and the possibility to worship. This is also the case of the Orthodox minority in Czechoslovakia, which was united and organized into an autocephalous church by the patriarchate of Moscow in 1951. Only in Albania did a Communist government announce the total liquidation of organized religion, following the Cultural Revolution of 1966–68.

Among the national Orthodox churches, the Church of Greece is the only one that preserved the legal status it acquired in the 19th century as the national state church. As such, it was supported by the successive political regimes of Greece. It could also develop an impressive internal mission. The Brotherhood Zoe ("Life"), organized according to the pattern of Western religious orders, was successful in creating a large system of church schools.

The Communist governments throughout eastern Europe collapsed during the late 1980s and early 1990s, effectively dissolving state control over churches and bringing new political and religious freedoms into the region.

The Orthodox Church in the Middle East. As a result of the Greco-Turkish War, the entire Greek population of Asia Minor was transferred to Greece (1922); the Orthodox

Conditions of the Eastern patriarchates

under the immediate jurisdiction of the ecumenical patriarchate of Constantinople were thus reduced to the Greek population of Istanbul and its vicinity. This population, rapidly shrinking in recent years, is now reduced to a few thousand. Still recognized as holding an honorary primacy among the Orthodox churches, the ecumenical patriarchate also exercises jurisdiction over several dioceses of the "diaspora" and, by consent of the Greek government, over the Greek islands. The impressive personality of Patriarch Athenagoras I (1948–72), who was succeeded by Dimitrios, contributed to its prestige on the pan-Orthodox and ecumenical levels. The patriarchate convened pan-Orthodox conferences in Rhodes, Belgrade, Geneva, and other cities and began preparations for a "Great Council" of the Orthodox Church.

Together with the ecumenical patriarchate, the ancient sees of Alexandria, Antioch, and Jerusalem are remnants of the Byzantine imperial past, but under the present conditions they still possess many opportunities of development: Alexandria, as the centre of emerging African communities (see below *The Orthodox diaspora and missions*); Antioch, as the largest Arab Christian group, with dioceses in Syria, Lebanon, and Iraq; and Jerusalem, as the main custodian of the Christian holy places in that city.

The two ancient churches of Cyprus and Georgia, with their quite peculiar history, continue to play important roles among the Orthodox sister churches. Autocephalous since 431, the Church of Cyprus survived the successive occupations, and often oppressions, by the Arabs, the Crusaders, the Venetians, the Turks, and the English. Following the pattern of all areas where Islām was predominant, the archbishop is traditionally seen as the ethnarch of the Greek Christian Cypriots. Archbishop Makarios also became the first president of the independent Republic of Cyprus in 1960. The Church of Georgia, isolated in the Caucasus in a country that became part of the Russian Empire in 1801, is the witness of one of the most ancient Christian traditions. It received autocephaly from its mother Church of Antioch as early as the 6th century and developed a literary and artistic civilization in its own language. Its head bears the traditional title of "Catholicos-Patriarch." When the Russians annexed the country in 1801, they suppressed Georgia's autocephaly and the church was governed by a Russian "exarch" until 1917 when the Georgians reestablished their ecclesiastical independence. Fiercely persecuted during the 1920s, the Georgian Church survives to the present day as an autocephalous patriarchate.

Orthodoxy in the United States. The first Orthodox communities in what is today the continental United States were established in Alaska and on the West Coast, as the extreme end of the Russian missionary expansion through Siberia (see above *The church in imperial Russia*). Russian monks settled on Kodiak Island in 1794. Among them was St. Herman (died 1837, canonized 1970), an ascetic and a defender of the natives' rights against their exploitation by ruthless Russian traders. After the sale of Alaska to the United States, a separate diocese "of the Aleutian Islands and Alaska" was created by the Holy Synod (1870). After the transfer of the diocesan centre to San Francisco and its renaming as the diocese "of the Aleutian Islands and North America" (1900), the original church establishment exercised its jurisdiction on the entire North American continent. In the 1880s, it accepted back into Orthodoxy hundreds of "Uniate" parishes of immigrants from Galicia and Carpatho-Russia, particularly numerous in the northern industrial states and in Canada. It also served the needs of immigrants from Serbia, Greece, Syria, Albania, and other countries. Some Greek and Romanian communities, however, invited priests directly from the mother country without official contact with the American bishop.

Organization and canonical status of American Orthodox churches

In 1905 the American archbishop Tikhon (future patriarch of Moscow) presented to the Russian synod the project of an autonomous, or autocephalous, church of America, whose structure would reflect the ethnic pluralism of its membership. He also foresaw the inevitable Americanization of his flock and encouraged the translation of the liturgy into English.

These projects, however, were hampered by the tragedies that befell the Russian Church following the Russian Revolution. The administrative system of the Russian Church collapsed. The non-Russian groups of immigrants sought and obtained their affiliation with mother churches abroad. In 1921 a "Greek Archdiocese of North and South America" was established by the ecumenical patriarch Meletios IV Metaxakis. Further divisions within each national group occurred repeatedly, and several independent jurisdictions added to the confusion.

A reaction against this chaotic pluralism manifested itself in the 1950s. More cooperation between the jurisdictions and a more systematic theological education contributed to an increased desire for unity. A Standing Conference of Canonical Orthodox Bishops in the Americas was established in 1960. In 1970 the patriarch of Moscow, reviving Tikhon's project of 1905, formally proclaimed its diocese in America (which had been in conflict with Moscow since 1931 on the issue of "loyalty" to the Soviet Union) as the autocephalous Orthodox Church in America, totally independent from administrative connections abroad. The ecumenical patriarchate of Constantinople, however, protested this move, turned down a request for autonomy presented by the Greek archdiocese (the largest single Orthodox body in the United States), and reiterated its opposition to the use of English in the liturgy (1970). This latest crisis of American Orthodoxy involves the very understanding of the Orthodox presence in the Western world, centring on the question of the utility of preserving the ethnic ties of the past.

The Orthodox diaspora and missions. Since World War I, millions of east Europeans were dispersed in various areas where Orthodox communities had never existed before. The Russian Revolution provoked a massive political emigration, predominantly to western Europe and particularly France. It included eminent churchmen, theologians, and Christian intellectuals, such as Bulgakov, Berdyayev, and V.V. Zenkovsky, who were able not only to establish in Paris a theological school of great repute but also to contribute significantly to the ecumenical movement. In 1922 Patriarch Tikhon appointed Metropolitan Evlogy to head the émigré churches, with residence in Paris. The authority of the metropolitan was challenged, however, by a group of bishops who had left their sees in Russia, retreating with the White armies, and who had found refuge in Sremski-Karlovci as guests of the Serbian Church. Despite several attempts at reconciliation, the "Synod" of Karlovci, proclaiming its firm attachment to the principle of tsarist monarchy, refused to recognize any measure taken by the reestablished patriarchate of Moscow. This group transferred its headquarters to New York and is also known as the "Russian Orthodox Church outside of Russia." It has no canonical relation with the official Orthodox patriarchates and churches. A "Ukrainian Orthodox Church in exile" finds itself in a similarly irregular canonical situation. Other émigré groups found refuge under the canonical auspices of the ecumenical patriarchate.

Theological and missionary activities after World War I

After World War II, a very numerous Greek emigration took place to western Europe, Australia, New Zealand, and Africa. In East Africa, without much initial effort on their part, these Greek-speaking emigrants have attracted a sizable number of black Christians, who have discovered in the Orthodox liturgy and sacramental worship a form of Christianity more acceptable to them than the more dogmatic institutions of Western Christianity. Also, in their eyes, Orthodoxy has the advantage of having no connection with the colonial regimes of the past. Orthodox communities, with an ever increasing number of native clergy, are spreading in Uganda, Kenya, and Tanzania. Less professionally planned than the former Russian missions in Alaska and Japan, these young churches constitute an interesting development in African Christianity.

Ecumenical involvement. Between the two world wars, many Orthodox churchmen of the ecumenical patriarchate of Constantinople, of Greece, of the Balkan churches, and of the Russian emigration took part in the ecumenical movement. After World War II, however, the churches of the Communist-dominated countries failed to join the newly created World Council of Churches (1948); only Constantinople and Greece did so. The situation changed

drastically in 1961, when the patriarchate of Moscow applied for membership and was soon followed by other autocephalous churches. Before and after 1961, the Orthodox consistently declared that their membership did not imply any relativistic understanding of the Christian truth, but that they were ready to discuss with all Christians the best way of restoring the lost unity of Christendom, as well as problems of common Christian action and witness in the modern world.

During the reign of Pope John XXIII, when Roman Catholicism became actively involved in ecumenism, the Orthodox—after some hesitation—contributed to the new atmosphere. The spectacular meetings in the 1960s between Patriarch Athenagoras and the Pope in Jerusalem, Istanbul, and Rome, the symbolic lifting of ancient anathemas, and other gestures were signs of rapprochement, although they are sometimes mistakenly interpreted as if they were ending the Schism itself. In the Orthodox view, full unity can be restored only in the fullness of truth witnessed by the entire church and sanctioned in sacramental communion.

Doctrine

COUNCILS AND CONFESSIONS

All Orthodox credal formulas, liturgical texts, and doctrinal statements affirm the claim that the Orthodox Church has preserved the original apostolic faith, which was also expressed in the common Christian tradition of the first centuries. The Orthodox Church recognizes as ecumenical the seven councils of Nicaea I (325), Constantinople I (381), Ephesus (431), Chalcedon (451), Constantinople II (553), Constantinople III (681), and Nicaea II (787) but considers that the decrees of several other later councils also reflect the same original faith (*e.g.,* the councils of Constantinople that endorsed the theology of St. Gregory Palamas in the 14th century). Finally, it recognizes itself as the bearer of an uninterrupted living tradition of true Christianity that is expressed in its worship, in the lives of the saints, and in the faith of the whole people of God.

The first seven ecumenical councils

In the 17th century, as a counterpart to the various "confessions" of the Reformation, there appeared several "Orthodox confessions," endorsed by local councils but, in fact, associated with individual authors (*e.g.,* Metrophanes Critopoulos, 1625; Peter Mogila, 1638; Dositheos of Jerusalem, 1672). None of these confessions would be recognized today as having anything but historical importance. When expressing the beliefs of his church, the Orthodox theologian, rather than seeking literal conformity with any of these particular confessions, will rather look for consistency with Scripture and tradition, as it has been expressed in the ancient councils, the early Fathers, and the uninterrupted life of the liturgy. He will not shy away from new formulations if consistency and continuity of tradition are preserved.

What is particularly characteristic of this attitude toward the faith is the absence of any great concern for establishing external *criteria* of truth—a concern that has dominated Western Christian thought since the Middle Ages. Truth appears as a living experience accessible in the communion of the church and of which the Scriptures, the councils, and theology are the normal expressions. Even ecumenical councils, in the Orthodox perspective, need subsequent "reception" by the body of the church in order to be recognized as truly ecumenical. Ultimately, therefore, truth is viewed as its own criterion: there are signs that point to it, but none of these signs is a substitute for a free and personal experience of truth, which is made accessible in the sacramental fellowship of the church.

The criterion of truth

Because of this view of truth, the Orthodox have traditionally been reluctant to involve church authority in defining matters of faith with too much precision and detail. This reluctance is not due to relativism or indifference but rather to the belief that truth needs no definition to be the object of experience and that legitimate definition, when it occurs, should aim mainly at excluding error and not at pretending to reveal the truth itself that is believed to be ever present in the church.

GOD AND MAN

The development of the doctrines concerning the Trinity and the incarnation, as it took place during the first eight centuries of Christian history, was related to the concept of man's *participation* in divine life.

The Greek Fathers of the church always implied that the phrase found in the biblical story of the creation of man (Gen. 1:26), according to "the image and likeness of God," meant that man is not an autonomous being and that his ultimate nature is defined by his relation to God, his "prototype." In paradise Adam and Eve were called to participate in God's life and to find in him the natural growth of their humanity "from glory to glory." To be "in God" is, therefore, the *natural* state of man. This doctrine is particularly important in connection with the Fathers' view of human freedom. For theologians such as Gregory of Nyssa (4th century) and Maximus the Confessor (7th century) man is truly free only when he is in communion with God; otherwise he is only a slave to his body or to "the world," over which, originally and by God's command, he was destined to rule.

Human freedom and sin

Thus, the concept of *sin* implies separation from God and the reduction of man to a separate and autonomous existence, in which he is deprived of both his natural glory and his freedom. He becomes an element subject to cosmic determinism, and the image of God is thus blurred within him.

Freedom in God, as enjoyed by Adam, implied the possibility of falling away from God. This is the unfortunate choice made by man, which led Adam to a subhuman and unnatural existence. The most unnatural aspect of his new state was death. In this perspective, "original sin" is understood not so much as a state of guilt inherited from Adam but as an unnatural condition of human life that ends in death. Mortality is what each man now inherits at his birth and this is what leads him to struggle for existence, to self-affirmation at the expense of others, and ultimately to subjection to the laws of animal life. The "prince of this world" (*i.e.,* Satan), who is also the "murderer from the beginning," has dominion over man. From this vicious circle of death and sin, man is understood to be liberated by the death and Resurrection of Christ, which is actualized in Baptism and the sacramental life in the church.

The general framework of this understanding of the God–man relationship is clearly different from the view that became dominant in the Christian West—*i.e.,* the view that conceived of "nature" as distinct from "grace" and that understood original sin as an inherited guilt rather than as a deprivation of freedom. In the East, man is regarded as fully man when he participates in God; in the West, man's nature is believed to be autonomous, sin is viewed as a punishable crime, and grace is understood to grant forgiveness. Hence, in the West, the aim of the Christian is justification, but in the East, it is rather communion with God and deification. In the West, the church is viewed in terms of mediation (for the bestowing of grace) and authority (for guaranteeing security in doctrine); in the East, the church is regarded as a communion in which God and man meet once again and a personal experience of divine life becomes possible.

Deification

CHRIST

The Orthodox Church is formally committed to the Christology (doctrine of Christ) that was defined by the councils of the first eight centuries. Together with the Latin Church of the West, it has rejected Arianism (a belief in the subordination of the Son to the Father) at Nicaea (325), Nestorianism (a belief that stresses the independence of the divine and human natures of Christ) at Ephesus (431), and Monophysitism (a belief that Christ had only one divine nature) at Chalcedon (451). The Eastern and Western churches still formally share the tradition of subsequent Christological developments, even though the famous formula of Chalcedon, "one person in two natures," is given different emphases in the East and West. The stress on Christ's identity with the preexistent Son of God, the Logos (Word) of the Gospel According to John, characterizes Orthodox Christology. On Byzantine icons, around the

Logos as the Christ

face of Jesus, the Greek letters ὁ ὢν—the equivalent of the Jewish Tetragrammaton YHWH, the name of God in the Old Testament—are often depicted. Jesus is thus always seen in his divine identity. Similarly, the liturgy consistently addresses the Virgin Mary as Theotokos (the "one who gave birth to God"), and this term, formally admitted as a criterion of orthodoxy at Ephesus, is actually the only "Mariological" (doctrine of Mary) dogma accepted in the Orthodox Church. It reflects the doctrine of Christ's unique divine Person, and Mary is thus venerated only because she is his mother "according to the flesh."

This emphasis on the personal divine identity of Christ, based on the doctrine of St. Cyril of Alexandria (5th century), does not imply the denial of his humanity. The anthropology (doctrine of man) of the Eastern Fathers does not view man as an autonomous being but rather implies that communion with God makes man fully human. Thus the human nature of Jesus Christ, fully assumed by the divine Word, is indeed the "new Adam" in whom the whole of humanity receives again its original glory. Christ's humanity is fully "ours"; it possessed all the characteristics of the human being—"each nature (of Christ) acts according to its properties," Chalcedon proclaimed, following Pope Leo—without separating itself from the divine Word. Thus, in death itself—for Jesus' death was indeed a fully human death—the Son of God was the "subject" of the Passion. The theopaschite formula ("God suffered in the flesh") became, together with the Theotokos formula, a standard of orthodoxy in the Eastern Church, especially after the second Council of Constantinople (553). It implied that Christ's humanity was indeed real not only in itself but also for God, since it brought him to death on the cross, and that the salvation and redemption of humanity can be accomplished by God alone—hence the necessity for him to condescend to death, which held humanity captive.

Theology of salvation and redemption

This theology of redemption and salvation is best expressed in the Byzantine liturgical hymns of Holy Week and Easter: Christ is the one who "tramples down death by death," and, on the evening of Good Friday, the hymns already exalt his victory. Salvation is conceived not in terms of satisfaction of divine justice, through paying the debt for the sin of Adam—as the medieval West understood it—but in terms of uniting the human and the divine with the divine overcoming human mortality and weakness and, finally, exalting man to divine life.

What Christ accomplished once and for all must be appropriated freely by those who are "in Christ"; their goal is "deification," which does not mean dehumanization but the exaltation of man to the dignity prepared for him at creation. Such feasts as the Transfiguration or the Ascension are extremely popular in the East precisely because they celebrate humanity glorified in Christ—a glorification that anticipates the coming of the Kingdom of God, when God will be "all in all."

Participation in the already deified humanity of Christ is the true goal of Christian life, and it is accomplished through the Holy Spirit.

THE HOLY SPIRIT

The gift of the Holy Spirit at Pentecost "called all men into unity," according to the Byzantine liturgical hymn of the day; into this new unity, which St. Paul called the "body of Christ," each individual Christian enters through Baptism and "chrismation" (the Eastern form of the Western "confirmation") when the priest anoints him saying "the seal of the gift of the Holy Spirit."

This gift, however, requires man's free response. Orthodox saints such as Seraphim of Sarov (died 1833) described the entire content of Christian life as a "collection of the Holy Spirit." The Holy Spirit is thus conceived as the main agent of man's restoration to his original natural state through Communion in Christ's body. This role of the Spirit is reflected, very richly, in a variety of liturgical and sacramental acts. Every act of worship usually starts with a prayer addressed to the Spirit, and all major sacraments begin with an invocation to the Spirit. The eucharistic liturgies of the East attribute the ultimate mystery of Christ's Presence to a descent of the Spirit upon the worshipping congregation and upon the eucharistic bread and wine. The significance of this invocation (in Greek *epiklésis*) was violently debated between Greek and Latin Christians in the Middle Ages because the Roman canon of the mass lacked any reference to the Spirit and was thus considered as deficient by the Orthodox Greeks.

The Holy Spirit as the agent of restoration

Since the Council of Constantinople (381), which condemned the Pneumatomachians ("fighters against the Spirit"), no one in the Orthodox East has ever denied that the Spirit is not only a "gift" but also the giver—*i.e.*, that he is the third Person of the holy Trinity. The Greek Fathers saw in Gen. 1:2 a reference to the Spirit's cooperation in the divine act of creation; the Spirit was also viewed as active in the "new creation" that occurred in the womb of the Virgin Mary when she became the mother of Christ (Luke 1:35); and finally, Pentecost was understood to be an anticipation of the "last days" (Acts 2:17) when, at the end of history, a universal communion with God will be achieved. Thus, all the decisive acts of God are accomplished "by the Father in the Son, through the Holy Spirit."

THE HOLY TRINITY

By the 4th century a polarity developed between the Eastern and Western Christians in their respective understandings of the Trinity. In the West God was understood primarily in terms of one essence (the Trinity of Persons being conceived as an irrational truth found in revelation); in the East the tri-personality of God was understood as the *primary* fact of Christian experience. For most of the Greek Fathers, it was not the Trinity that needed theological proof but rather God's essential unity. The Cappadocian Fathers (Gregory of Nyssa, Gregory of Nazianzus, and Basil of Caesarea) were even accused of being tri-theists because of the personalistic emphasis of their conception of God as one essence in three hypostases (the Greek term *hypostasis* was the equivalent of the Latin *substantia* and designated a concrete reality). For Greek theologians, this terminology was intended to designate the concrete New Testamental revelation of the Son and the Spirit, as distinct from the Father.

Modern Orthodox theologians tend to emphasize this personalistic approach to God; they claim that they discover in it the original biblical personalism, unadulterated in its content by later philosophical speculation.

The *Filioque* controversy

Polarization of the Eastern and the Western concepts of the Trinity is at the root of the *Filioque* dispute. The Latin word *Filioque* ("and from the Son") was added to the Nicene Creed in Spain in the 6th century. By affirming that the Holy Spirit proceeds not only "from the Father" (as the original creed proclaimed) but also "from the Son," the Spanish councils intended to condemn Arianism by reaffirming the Son's divinity. Later, however, the addition became an anti-Greek battle cry, especially after Charlemagne (9th century) made his claim to rule the revived Roman Empire. The addition was finally accepted in Rome under German pressure. It found justification in the framework of Western conceptions of the Trinity; the Father and the Son were viewed as one God in the act of "spiration" of the Spirit.

The Byzantine theologians opposed the addition, first on the ground that the Western Church had no right to change the text of an ecumenical creed unilaterally and, second, because the *Filioque* clause implied the reduction of the divine persons to mere relations ("the Father and the Son are two in relation to each other, but one in relation to the Spirit"). For the Greeks the Father alone is the origin of both the Son and the Spirit. Patriarch Photius (9th century) was the first Orthodox theologian to explicitly spell out the Greek opposition to the *Filioque* concept, but the debate continued throughout the Middle Ages.

THE TRANSCENDENCE OF GOD

Apophatic theology

An important element in the Eastern Christian understanding of God is the notion that God, in his essence, is totally transcendent and unknowable and that, strictly speaking, God can only be designated by negative attributes: it is possible to say what God is not, but it is impossible to say what he is.

A purely negative, or "apophatic" theology—the only one applicable to the *essence* of God in the Orthodox view—does not lead to agnosticism, however, because God reveals himself personally—as Father, Son, and Holy Spirit—and also in his *acts*, or "energies." Thus, true knowledge of God always includes three elements: religious awe; personal encounter; and participation in the acts, or energies, which God freely bestows on creation.

This conception of God is connected with the personalistic understanding of the Trinity. It also led to the official confirmation by the Orthodox Church of the theology of St. Gregory Palamas, the leader of Byzantine hesychasts (monks devoted to divine quietness through prayer), at the councils of 1341 and 1351 in Constantinople. The councils confirmed a real distinction in God, between the unknowable essence and the acts, or "energies," which make possible a real communion with God. The deification of man, realized in Christ once and for all, is thus accomplished by a communion of divine energy with humanity in Christ's glorified manhood.

MODERN THEOLOGICAL DEVELOPMENTS

Until the conquest of Constantinople by the Turks (1453), Byzantium was the unquestioned intellectual centre of the Orthodox Church. Far from being monolithic, Byzantine theological thought was often polarized by a Humanistic trend, favouring the use of Greek philosophy in theological thinking, and the more austere and mystical theology of the monastic circles. The concern for preservation of Greek culture and for the the political salvation of the empire led several prominent Humanists to adopt a position favourable to union with the West. The most creative theologians (*e.g.*, Symeon the New Theologian, died 1033; Gregory Palamas, died 1359; Nicholas Cabasilas, died *c.* 1390), however, were found rather in the monastic party that continued the tradition of patristic spirituality based upon the theology of deification.

The dark age of Orthodox theology

The 16th, 17th, and 18th centuries were the dark age of Orthodox theology. Neither in the Middle East nor in the Balkans nor in Russia was there any opportunity for independent theological creativity. Since no formal theological education was accessible, except in Western Roman Catholic or Protestant schools, the Orthodox tradition was preserved primarily through the liturgy, which retained all its richness and often served as a valid substitute for formal schooling. Most doctrinal statements of this period, issued by councils or by individual theologians, were polemical documents directed against Western missionaries.

After the reforms of Peter the Great (died 1725), a theological school system was organized in Russia. Shaped originally in accordance with Western Latin models and staffed with Jesuit-trained Ukrainian personnel, this system developed, in the 19th century, into a fully independent and powerful tool of theological education. The Russian theological efflorescence of the 19th and 20th centuries produced many scholars, especially in the historical field (*e.g.*, Philaret Drozdov, died 1867; V.O. Klyuchevsky, died 1913; V.V. Bolotov, died 1900; E.E. Golubinsky, died 1912; N.N. Glubokovsky, died 1937). Independently of the official theological schools, a number of laymen with secular training developed theological and philosophical traditions of their own and exercised a great influence on modern Orthodox theology (*e.g.*, A.S. Khomyakov, died 1860; V.S. Solovyev, died 1900; N. Berdyayev, died 1948), and some became priests (P. Florensky, died 1943; S. Bulgakov, died 1944). A large number of the Russian theological intelligentsia (*e.g.*, S. Bulgakov, G. Florovsky) emigrated to western Europe after the Russian Revolution (1917) and played a leading role in the ecumenical movement.

With the independence of the Balkans, theological schools were also created in Greece, Serbia, Bulgaria, and Romania. Modern Greek scholars contributed to the publication of important Byzantine ecclesiastical texts and produced standard theological textbooks.

The Orthodox diaspora

The Orthodox diaspora—the emigration from eastern Europe and the Middle East—in the 20th century has contributed to modern theological development through their establishment of theological centres in western Europe and America.

Orthodox theologians reacted negatively to the new dogmas proclaimed by Pope Pius IX: the Immaculate Conception of Mary (1854) and papal infallibility (1870). In connection with the dogma of the Assumption of Mary, proclaimed by Pope Pius XII (1950), the objections mainly concerned the presentation of such a tradition in the form of a dogma.

In contrast to the recent general trend of Western Christian thought toward social concerns, Orthodox theologians generally emphasize that the Christian faith is primarily a direct experience of the Kingdom of God, sacramentally present in the church. Without denying that Christians have a social responsibility to the world, they consider this responsibility as an outcome of the life in Christ. This traditional position accounts for the remarkable survival of the Orthodox Churches under the most contradictory and unfavourable of social conditions, but, to Western eyes, it often appears as a form of passive fatalism.

The structure of the church

THE CANONS

The permanent criteria of church structure for the Orthodox Church today, outside of the New Testament writings, are found in the canons (regulations and decrees) of the first seven ecumenical councils; the canons of several local or provincial councils, whose authority was recognized by the whole church; the so-called *Apostolic Canons* (actually some regulations of the church in Syria, dating from the 4th century); and the "canons of the Fathers," or selected extracts from prominent church leaders having canonical importance.

The nomocanon

A collection of these texts was made in the Byzantine nomocanon, attributed, in its final form, to the patriarch Photius (9th century). The Byzantine Church, as well as the modern Orthodox Churches, has adapted the general principles of this collection to its particular situation, and the local autocephalous churches govern themselves according to their own particular statutes, although all accept the ancient canons as their common canonical reference.

The canons themselves do not represent a system or a code. They do, however, reflect a consistent view of the church, of its mission, and of its various ministries; they also reflect an evolution of ecclesiastical structure—*i.e.*, the growth of centralization in the framework of the Christian Roman Empire. For the Orthodox Church today, only the original self-understanding of the church has a theologically normative value. Thus, those canons that reflect the nature of the church as the body of Christ have an unchanging validity today; other canons, if they can be recognized as conditioned by the historical situation in which they were issued, are subject to change by conciliar authority; others have simply fallen out of practice. The use and interpretation of the canons is therefore possible only in the light of some understanding of the church's nature. This theological dimension is the ultimate criterion through which it is possible to distinguish what is permanent in the canons from that which represents no more than a historical value.

THE EPISCOPATE

The Orthodox understanding of the church is based on the principle, attested to in the canons and in early Christian tradition, that each local community of Christians, gathered around its bishop and celebrating the Eucharist, is the local realization of the *whole* body of Christ. "Where Christ is, there is the Catholic church," wrote Ignatius of Antioch (*c.* AD 100). Modern Orthodox theology also emphasizes that the office of the bishop is the highest among the sacramental ministries and that there is therefore no divinely established authority *over* that of the bishop in his own community, or diocese. Neither the local churches nor the bishops, however, can or should live in isolation. The wholeness of church life, realized in each local community, is regarded as identical with that of the other local churches in the present and in the past. This identity and continuity is manifested in the act of the ordination of bishops, an act that requires the presence of several other

bishops in order to constitute a conciliar act and to witness to the continuity of apostolic succession and tradition.

The bishop is primarily the guardian of the faith and, as such, the centre of the sacramental life of the community. The Orthodox Church maintains the doctrine of apostolic succession—*i.e.*, the idea that the ministry of the bishop must be in direct continuity with that of the Apostles of Jesus. Orthodox tradition—as expressed especially in its medieval opposition to the Roman papacy—distinguishes the office of the "Apostle" from that of the bishop, however, in that the first is viewed as a universal witness to the historic Jesus and his Resurrection, while the latter is understood in terms of the pastoral and sacramental responsibility for a local community, or church. The continuity between the two is, therefore, a continuity in faith rather than in function. This Orthodox concept of the doctrine of apostolic succession has received wider exposure in Western churches recently because of increased encounters and consultations between Orthodox and Anglican churchmen, the Orthodox always emphasizing unity of faith as a prerequisite for recognition, on their part, of the "validity" of Anglican orders.

Apostolic succession

No bishop can be consecrated or exercise his ministry without being in unity with his colleagues—*i.e.*, be a member of an episcopal council, or "synod." After the Council of Nicaea (325), whose canons are still effective in the Orthodox Church, each province of the Roman Empire had its own synod of bishops that acted as a fully independent unit for the consecration of new bishops and also as a high ecclesiastical tribunal. In the contemporary Orthodox Church these functions are fulfilled by the synod of each autocephalous church. In the early church the bishop of the provincial capital acted as chairman of the synod and was generally called "metropolitan." Today this function is fulfilled by the local primate who is sometimes called "patriarch" (in the autocephalous churches of Constantinople, Alexandria, Antioch, Jerusalem, Russia, Georgia, Serbia, Romania, and Bulgaria), but he may also carry the title of archbishop (Cyprus, Greece) or metropolitan (Poland, the Czech and Slovak republics, America). The titles of archbishop and metropolitan are also widely used as honorific distinctions.

Generally, but not always, the jurisdiction of each autocephalous synod coincides with national borders—the exceptions are numerous in the Middle East (*e.g.*, jurisdiction of Constantinople over the Greek islands, jurisdiction of Antioch over several Arab states, etc.)—and concerns also the national dioceses of the Orthodox diaspora (*e.g.*, western Europe, Australia, America), which frequently remain under the authority of their mother churches. The latter situation led to an uncanonical overlapping of Orthodox jurisdictions, all based on ethnic origins. Several factors, going back to the Middle Ages, have contributed to modern ecclesiastical nationalism in the Orthodox Church. These factors include the use of the vernacular in the liturgy and the subsequent identification of religion with national culture; this identification sometimes helps the survival of the church under adverse political conditions, but it also hampers missionary expansion and the sense of a specifically Christian identity of the faithful.

CLERGY AND LAITY

The emphasis on communion and fellowship, as the basic principle of church life, inhibited the development of clericalism. The early Christian practice of having the laity participate in episcopal elections never disappeared completely in the East. In modern times, it has been restored in several churches. The Moscow Council of 1917–1918 introduced it in Russia, even if the events of the Revolution prevented its full implementation. Bishops are also elected by clergy-laity conventions in America and in other areas of the Orthodox world.

Clerical marriage

The lower orders of the clergy—*i.e.*, priests and deacons—are generally married men. The present canonical legislation allows the ordination of married men to the diaconate and the priesthood, provided that they were married only once and that their wives are neither widows nor divorcees. These stipulations reflect the general principle of absolute monogamy, which the Eastern Church considered as a Christian norm to which candidates for the priesthood are to comply strictly. Deacons and priests cannot marry after their ordination.

Bishops, however, are selected from among the unmarried clergy or widowed priests. The rule defining the requirement for an unmarried episcopate was issued at a time (6th century) when monks represented the elite of the clergy. The contemporary decrease in the number of monks in the Orthodox Church has created a serious problem in some territorial churches, in that new candidates for the episcopacy are difficult to find.

Besides being admitted, at least in some areas, to participation in episcopal elections, Orthodox laymen often occupy positions in church administration and in theological education. In Greece almost all professional theologians are laymen. Laymen also frequently serve as preachers.

MONASTICISM

The tradition of Eastern Christian monasticism goes back to the 3rd and 4th centuries of the Christian Era. From its beginning it was essentially a contemplative movement seeking the experience of God in a life of permanent prayer. This contemplative character has remained its essential feature throughout the centuries. Eastern Christianity never experienced the development of religious orders, pursuing particular missionary or educational goals and organized on a universal scale, as did Western Christianity.

Concern for prayer, as the central and principal function of monasticism, does not mean that the Eastern Christian monastic movement was of a single uniform character. Eremitic (solitary) monasticism, favouring the personal and individual practice of prayer and asceticism, often competed with "cenobitic" (communal) monastic life, in which prayer was mainly liturgical and corporate. The two forms of monasticism originated in Egypt and coexisted in Byzantium, as well as throughout eastern Europe.

Eremitic and cenobitic monasticism

In Byzantium the great monastery of Studion became the model of numerous cenobitic communities (see above under *History: The church of imperial Byzantium*). It is in the framework of the eremitic, or Hesychast, tradition, however, that the most noted Byzantine mystical theologians (*e.g.*, Symeon the New Theologian, Gregory Palamas, etc.) received their training. One of the major characteristics of the Hesychast tradition is the practice of the "Jesus prayer," or constant invocation of the name of Jesus, sometimes in connection with breathing. This practice won wide acceptance in medieval and modern Russia.

Cenobitic traditions of Byzantium also were important in Slavic lands. The colonization of the Russian north was largely accomplished by monks who acted as pioneers of civilization and as missionaries.

In Byzantium, as well as in other areas of the Orthodox world, the monks were often the only upholders of the moral and spiritual integrity of Christianity, and thus they gained the respect of the masses, as well as that of the intellectuals. The famous Russian *startsy* ("elders") of the 19th century became the spiritual leaders of Dostoyevsky, Gogol, and Tolstoy and inspired many religious philosophers in their quest for religious experience.

Today the most famous, though declining, centre of Orthodox monasticism is Mt. Athos (Greece), where over a thousand monks of different national backgrounds form a variety of communities, grouped into a monastic republic.

Worship and sacraments

THE ROLE OF THE LITURGY

By its theological richness, spiritual significance, and variety, the worship of the Orthodox Church represents one of the most significant factors in this church's continuity and identity. It helps to account for the survival of Christianity during the many centuries of Muslim rule in the Middle East and the Balkans when the liturgy was the only source of religious knowledge or experience. Since liturgical practice was practically the only religious expression legally authorized in the former Soviet Union, the continuous existence of Orthodox communities in the region was also centred almost exclusively around the liturgy.

The concept that the church is most authentically itself

Worship as the primary experience of Eastern Christianity

when the congregation of the faithful is gathered together in worship is a basic expression of Eastern Christian experience. Without that concept it is impossible to understand the fundamentals of church structure in Orthodoxy, with the bishop functioning in his essential roles of teacher and high priest in the liturgy. Similarly, the personal experience of man's participation in divine life is understood in the framework of the continuous liturgical action of the community.

According to many authorities, one of the reasons that helps to explain why the Eastern liturgy has made a stronger impact on the Christian Church than has its Western counterpart is that it has always been viewed as a total experience, appealing simultaneously to the emotional, intellectual, and aesthetic faculties of man. The liturgy includes a variety of models, or symbols, using formal theological statements as well as bodily perceptions and gestures (*e.g.,* music, incense, prostrations) or the visual arts. All are meant to convey the content of the Christian faith to the educated and the noneducated alike. Participation in the liturgy implies familiarity with its models, and many of them are conditioned by the historical and cultural past of the church. Thus, the use of such an elaborate and ancient liturgy presupposes catechetical preparation. It may require an updating of the liturgical forms themselves. The Orthodox Church recognizes that liturgical forms are changeable and that, since the early church admitted a variety of liturgical traditions, such a variety is also possible today. Thus, Orthodox communities with Western rites now exist in western Europe and in the Americas.

The Orthodox Church, however, has always been conservative in liturgical matters. This conservatism is due, in particular, to the absence of a central ecclesiastical authority that could enforce reforms and to the firm conviction of the church membership as a whole that the liturgy is the main vehicle and experience of true Christian beliefs. Consequently, reform of the liturgy is often considered as equivalent to a reform of the faith itself. However inconvenient this conservatism may be, the Orthodox liturgy has preserved many essential Christian values transmitted directly from the experience of the early church.

Throughout the centuries, the Orthodox liturgy has been richly embellished with cycles of hymns from a wide variety of sources. Byzantium (where the present Orthodox liturgical rite took shape), while keeping many biblical and early Christian elements, used the lavish resources of patristic theology and Greek poetry, as well as some gestures of imperial court ceremonial, in order to convey the realities of God's kingdom.

Normally, the content of the liturgy is directly accessible to the faithful, because the Byzantine tradition is committed to the use of any vernacular language in the liturgy. Translation of both Scriptures and liturgy into various languages was undertaken by the medieval Byzantines, as well as by modern Russian missionaries. Liturgical conservatism, however, leads de facto to the preservation of antiquated languages. The Byzantine Greek used in church services by the modern Greeks and the Old Slavonic still preserved by all the Slavs are at least as distant from the spoken languages as is the language of the King James Version—used in many Protestant Churches—from modern English.

THE EUCHARISTIC LITURGIES

Two eucharistic liturgies are most generally used in Orthodox worship—*i.e.,* the so-called liturgies of St. John Chrysostom and of St. Basil the Great. Both acquired their present shape by the 9th century, but it is generally recognized that the wording of the eucharistic "canon" of the liturgy of St. Basil goes back to the 4th century—*i.e.,* to St. Basil himself. The so-called Liturgy of St. James is used occasionally, especially in Jerusalem. During the period of Lent, a service of Communion, with elements (bread and wine) reserved from those consecrated on the previous Sunday, is celebrated in connection with the evening service of Vespers; it is called the "Liturgy of the Presanctified" and is attributed to St. Gregory the Great.

The liturgies of St. John Chrysostom and of St. Basil differ only in the text of the eucharistic canon: their overall structures, established in the high Middle Ages, are identical.

These eucharistic liturgies begin with an elaborate rite of preparation (*proskomidē*). A priest on a separate "table of oblation" disposes on a paten (plate) the particles of bread that will symbolize the assembly of the saints, both living and dead, around Christ, the "Lamb of God." Then follows the "Liturgy of the Catechumens," which begins with a processional entrance of the priest into the sanctuary with the Gospel (Little Entrance) and which includes the traditional Christian "liturgy of the word"—*i.e.,* the reading from the New Testament letters and the Gospels as well as a sermon. This part of the liturgy ends with the expulsion of the "catechumens," who, until they were baptized, were not admitted to the sacramental part of the service. The "Liturgy of the Faithful" includes another ceremonial procession of the priest into the sanctuary. He carries the bread and wine from the table of oblations to the altar (Great Entrance); the following recitation of the Nicene Creed, the eucharistic canon, the Lord's Prayer, and Communion are—as in the West—the characteristic parts of the Byzantine "Liturgy of the Faithful." The bread used for the Eucharist is ordinary leavened bread; both elements (bread and wine) are distributed with a special spoon (*labis*).

THE LITURGICAL CYCLES

Variety of hymnodical texts

One of the major characteristics of the Byzantine liturgical tradition is the wealth and variety of hymnodical texts marking the various cycles of the liturgical year. A special liturgical book contains the hymns for each of the main cycles. The *daily* cycle includes the offices of Hesperinos (Vespers), Apodeipnon (Compline), the midnight prayer, Orthros (Matins), and the four canonical "hours"—*i.e.,* offices to be said at the "First" (6:00 AM), "Third" (9:00 AM), "Sixth" (12:00 noon), and "Ninth" (3:00 PM) hours. The liturgical book covering the daily cycle is called the *Hōrologion* ("The Book of Hours"). The *Paschal* (Easter) cycle is centred on the "Feast of Feasts"—*i.e.,* of Christ's Resurrection: it includes the period of Great Fast (Lent), preceded by three Sundays of preparation and the period of 50 days following Easter. The hymns of the Lenten period are found in the *Triōdion* (Three Odes), and those of the Easter season in the *Pentēkostarion* (called the "Flowery Triodion"). The weekly cycle is the continuation of the Resurrection cycle found in the *Triōdion* and the *Pentēkostarion:* each week following the Sunday after Pentecost (50 days after Easter) possesses its own musical tone, or mode, in accordance with which all the hymns of the week are sung. There are eight tones whose composition is traditionally attributed to St. John of Damascus (8th century). Each week is centred around Sunday, the day of Christ's Resurrection.

The Easter and weekly cycles clearly dominate all offices of the entire year and illustrate the absolute centrality of the Resurrection in the Eastern understanding of the Christian message. The date of Easter, set at the Council of Nicaea (325), is the first Sunday after the full moon following the spring equinox. Differences between the East and West in computing the date exist because the Orthodox Church uses the Julian calendar for establishing the date of the equinox (hence a delay of 13 days) and also because of the tradition that Easter must necessarily follow the Jewish Passover and must never precede it or coincide with it. The *yearly* cycle includes the hymns for each of the 366 days of the calendar year, with its feasts and daily commemoration of saints. They are found in the 12 volumes of the *Menaion* ("Book of Months").

The golden age of Byzantine hymnody

From the 6th to the 9th century the Byzantine Church experienced its golden age of creativity in the writing of hymns by outstanding poets such as John of Damascus. In more recent times hymn writing has generally followed the accepted patterns set by these authors but rarely has it reached the quality of its models. Since the Eastern Orthodox tradition bans instrumental music, or accompaniment, the singing is always a cappella, with only a few exceptions admitted by Westernized parishes in America. The idea behind the ban is based upon the practice of

worship in the New Testament; *i.e.*, only the natural aptitudes of the living congregation are viewed as capable of expressing praise that is worthy of God. In many Orthodox churches there is a wealth of new musical compositions for liturgical texts.

THE SACRAMENTS

Contemporary Orthodox catechisms and textbooks all affirm that the church recognizes seven *mystēria*, or "sacraments": Baptism, chrismation, Communion, holy orders, penance, anointing of the sick (the "extreme unction" of the medieval West), and marriage. Neither the liturgical book called *Euchologion* (prayer book), which contains the texts of the sacraments, nor the patristic tradition, however, formally limits the *number* of sacraments; they do not distinguish clearly between the "sacraments" and such acts as the blessing of water on Epiphany day or the burial service or the service for the tonsuring of a monk that in the West are called *sacramentalia*. In fact, no council recognized by the Orthodox Church ever defined the number of sacraments; it is only through the "Orthodox confessions" of the 17th century directed against the Reformation that the number seven has been generally accepted. The underlying sacramental theology of the Orthodox Church is based, however, on the notion that the ecclesiastical community is the unique *mystērion*, of which the various sacraments or *sacramentalia* are the normal expressions.

The Christian community as the *mystērion*

In the West, since the Scholastic period (Middle Ages) and, especially, since the Catholic Reformation (16th century), much emphasis has been placed on the vicarious juridical power of the minister to administer the sacraments validly. The Orthodox East, however, interprets each sacramental act as a prayer of the entire ecclesiastical community, led by the bishop or his representative, and also as a response of God, based upon Christ's promise to send the Holy Spirit upon the church. These two aspects of the sacrament exclude both magic and legalism; they imply that the Holy Spirit is given to free men and call for their responses. In the *mystērion* of the church, the participation of men in God is effected through their "cooperation" or "synergy"; to make this participation possible once more is the goal of the incarnation.

Baptism and confirmation. Baptism is normally performed by triple immersion as a sign of the death and Resurrection of Christ; thus, the rite appears essentially as a gift of new life. It is immediately followed by confirmation, performed by the priest who anoints the newly baptized Christian with "Holy Chrism" (oil) blessed by the bishop. Baptized and confirmed children are admitted to Holy Communion. By admitting children immediately after their Baptism to both confirmation and Communion, the Eastern Christian tradition maintains the positive meaning of Baptism—*i.e.*, as the beginning of a new life nourished by the Eucharist.

The Eucharist. There never has been, in the East, much speculation about the nature of the eucharistic mystery. Both canons presently in use (that of St. Basil and that of St. John Chrysostom) include the "words of institution" ("This is my Body . . .," "This is my Blood . . ."), which are traditionally considered in the West as the formula necessary for the validity of the sacrament. In the East, however, the culminating point of the prayer is not in the remembrance of Christ's act but in the invocation of the Holy Spirit, which immediately follows: "Send down Thy Holy Spirit upon us and upon the Gifts here spread forth, and make this bread to be the precious Body of Thy Christ. . . ." Thus, the central mystery of Christianity is seen as being performed by the prayer of the church and through an invocation of the Spirit. The nature of the mystery that occurs in the bread and wine is signified by the term *metabolē* ("sacramental change"). The Western term transubstantiation occurs only in some confessions of faith after the 17th century.

Orders. The Orthodox Church recognizes three major orders: the diaconate, the priesthood, and the episcopate (bishop), as well as the minor orders of the lectorate and the subdiaconate. All the ordinations are performed by a bishop and, normally, during the eucharistic liturgy. The consecration of a bishop requires the participation of at least two or three bishops, as well as an election by a canonical synod.

Penance. The sacrament of penance in the early church was a solemn and public act of reconciliation, through which an excommunicated sinner was readmitted into church membership. Historically it has evolved into a private act of confession through which every Christian's membership in the church is periodically renewed. In the Orthodox Church today there is a certain variety in both the practice and the rite of penance. In the churches of the Balkans and the Middle East, it fell into disuse during the four centuries of Turkish occupation but is gradually being restored today. In Greek-speaking churches only certain priests, especially appointed by the bishop, have the right to hear confessions. In Russia, on the contrary, confessions remained a standard practice that was generally required before communion. General or group confession, introduced by John of Kronshtadt, a Russian spiritual leader of the early 20th century, is also occasionally practiced. The rite of confession in the *Euchologion* retains the form of a prayer, or invocation, said by the priest for the remission of the penitent's sins. In the Slavic ritual a Latin-inspired and juridical form of personal absolution was introduced by Peter Mogila, metropolitan of Kiev (17th century). Confession, in Orthodox practice, is generally viewed as a form of spiritual healing rather than as a tribunal. The relative lack of legalism reflects the Eastern patristic approach to sin—*i.e.*, as an internal passion and as an enslavement. The external sinful acts—which alone can be legally tried—are only manifestations of man's internal disease.

Practice of confession

Anointing of the sick. Anointing of the sick is a form of healing by prayer. In the Greek Church it is performed annually in church for the benefit of the entire congregation on the evening of Holy Wednesday.

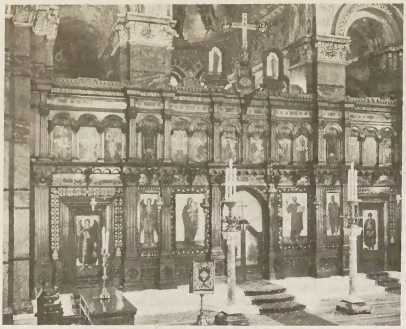
Marriage. Marriage is celebrated through a rite of crowning, performed with great solemnity and signifying an eternal union, sacramentally "projected" into the Kingdom of God. Orthodox theology of marriage insists on its sacramental eternity rather than its legal indissolubility. Thus, second marriages, in cases of either widowhood or divorce, are celebrated through a subdued penitential rite, and men who have been married more than once are not admitted to the priesthood. Remarriage after divorce is tolerated on the basis of the possibility that the sacrament of marriage was not originally received with the consciousness and responsibility that would have made it fully effective; according to this view, remarriage can be a second chance.

ARCHITECTURE AND ICONOGRAPHY

Since the time of Constantine I, Eastern Christianity has developed a variety of patterns in church architecture. The chief model was created when Emperor Justinian I completed the "great church" of Hagia Sophia in Constantinople in the 6th century. The architectural conception of that church consisted of erecting a huge round dome on top of the classical early Christian basilica. The dome was meant to symbolize the descent of heaven upon earth—*i.e.*, the ultimate meaning of the eucharistic celebration.

The screen, or iconostasis, which separates the sanctuary from the nave in contemporary Orthodox Churches, is a rather late development. After the triumph of orthodoxy over iconoclasm (destruction of images) in 843, a new emphasis was placed upon the permanent revelatory role of *images*. The incarnation implied that God had become man—*i.e.*, fully visible and thus describable in his human nature. The images of Christ and the saints, who had manifested in their lives the new humanity transfigured by the grace of God, were placed everywhere in full evidence before the congregation. A contrast was thus suggested between the visible manifestation of God through the pictural representation of Christ as man and his more perfect but mysterious and invisible presence in the Eucharist. The iconostasis, together with those parts of the liturgy that involve the closing and opening of the curtain before the altar, emphasizes the mysterious and "eschatological" (consummation of history) character of the eucharistic

The Iconoclastic Controversy



The iconostasis or screen, characteristically decorated with icon paintings, separates the sanctuary from the nave in an Orthodox church. In the Greek Orthodox Cathedral, London.
By courtesy of the Greek Orthodox Cathedral, London; photograph, A.C. Cooper Ltd.

service. They suggest, however, that this *mystery* is not a "secret" and that the Christian is being introduced through the eucharistic liturgy into the very reality of divine life and of the kingdom to come, which was revealed when God became man.

The long Iconoclastic Controversy (725–843), during which the Orthodox theology of icons was fully developed, concerned itself primarily with the problem of the incarnation: it was the direct continuation of the Christological debates of the 5th, 6th, and 7th centuries. The image of Christ, the incarnated God, became, for the Eastern Christian, a pictorial confession of faith: God was truly visible in the humanity of Jesus of Nazareth, and the saints—whose images surround that of Christ—are witnesses of the fact that the transfigured, "deified" humanity is accessible to those who believe in Christ.

Popular use of icons

Departing from tridimensional images or statues, which were reminiscent of pagan idolatry, the Christian East developed a rich tradition of iconography. Portable icons—often painted on wood but also using mosaics with enamel techniques—are always kept in houses or public places. Among the icon painters, who never signed their work, there appeared several artists of genius. Most of them are unknown, but tradition and written documents have revealed the names of some, such as the famous 14th–15th-century Russian painter Andrey Rublyov.

By courtesy of the State Tretyakov Gallery, Moscow



"The Old Testament Trinity," Russian icon painted by Andrey Rublyov, *c.* 1410. In the State Tretyakov Gallery, Moscow.

The church and the world

The schism between the Greek and the Latin churches coincided chronologically with a surge of Christian missionary activity in northern and eastern Europe. Both sides contributed to the resultant expansion of Christianity but used different methods. The West imposed a Latin liturgy on the new converts and thus made Latin the only vehicle of Christian civilization and a major instrument of ecclesiastical unity. The East, meanwhile, as noted above, accepted from the start the principle of translating both the Scriptures and the liturgy into the spoken tongues of the converted nations. Christianity thus became integrated into the indigenous cultures of the Slavic nations, and the universal Orthodox Church evolved as a fellowship of national churches rather than as a centralized body.

MISSIONS: ANCIENT AND MODERN

The Christian East, in spite of the integrating forces of Christian Hellenism, was always culturally pluralistic: since the first centuries of Christianity, Syrians, Armenians, Georgians, Copts, Ethiopians, and other ethnic groups used their own languages in worship and developed their own liturgical traditions. Even though by the time of the Greek missions to the Slavs the Byzantine Church was almost monolithically Greek, the idea of a liturgy in the vernacular was still quite alive, as is demonstrated by the use of the Slavic language by the missionaries of SS. Cyril and Methodius in the 9th century.

The Turkish conquest of the Middle East and of the Balkans (15th century) interrupted the missionary expansion of the Orthodox Church. Throughout the Middle Ages, Islām and Christianity had usually confronted each other only militarily, and the victory of Islām meant that the Christians could survive only in enclaves and were legally excluded from proselytizing among Muslims.

Russian missionary activity

The Russian Church alone was able to continue the tradition of SS. Cyril and Methodius, and it did so almost without interruption until the modern period. In the 14th century St. Stephen of Perm translated the Scriptures and the liturgy into the language of a Finnish tribe of the Russian north and became the first bishop of the Zyrians. The expansion of the Russian Empire in Asia was accompanied by efforts of evangelization that—sometimes in opposition to the avowed policy of Russianization practiced by the government of St. Petersburg—followed the Cyrillo-Methodian pattern of translation. This method was utilized among the Tatars of the Volga in the 16th century and among the various peoples of Siberia throughout the 18th and the 19th centuries. In 1714 a mission was established in China. In 1794 monks of the Valamo Abbey reached Alaska; their spiritual leader, the monk Herman, was canonized by the Orthodox Church in 1970. Missions in the Islāmic sphere resumed to the extent that by the year 1903 the liturgy was celebrated in more than 20 languages in the region of Kazan.

The Alaskan mission was under the direction of a modest priest sent to America from eastern Siberia, Ivan Veniaminov. During his long stay in America, first as a priest, then as a bishop (1824–68), he engaged in the work of translating the Gospels and the liturgy into the languages of the Aleuts, the Tlingit Indians, and the Eskimos of Alaska.

In Japan an Orthodox Church was established by the recently canonized St. Nikolay Kasatkin (died 1913). The distinctively Japanese character of this church enabled it to survive the political trials of the Russo-Japanese War (1904–05), of the Russian Revolution, and of World War II. The new Church of Japan received its full autonomy from the Russian Church in 1970.

The missionary tradition is also being revived in Greece.

Various Greek associations are dedicated to the pursuit of missionary work in East Africa, where sizable indigenous groups have recently joined the Orthodox Church.

ORTHODOXY AND OTHER CHRISTIANS

Since the failure of the unionist Council of Florence (1439) there have been no official attempts to restore unity between the Orthodox Church and Roman Catholicism. In 1484 an Orthodox council defined that Roman Catholics desiring to join the Orthodox Church were to be received through chrismation (or confirmation). In the 17th century, however, the relations deteriorated to the point that the ecumenical patriarchate of Constantinople decreed that all Roman Catholic and Protestant sacraments, including Baptism, were totally unauthentic. A parallel attitude prevailed in Russia until the 18th century, when large numbers of Eastern Rite Roman Catholics ("Uniates") were received back into Orthodoxy by a simple confession of faith, and this practice was adopted in the acceptance of individual Roman Catholics as well.

After the 16th-century Reformation, a lengthy correspondence took place between the Tübingen group of Reformers (German Lutheran) headed by P. Melanchthon and ecumenical Patriarch Jeremias II. It led to no concrete results, for the East generally considered the Protestants as only a branch of deviation of the altogether erroneous Roman Church.

Various attempts at rapprochement with the Anglican Communion, especially since the 19th century, were generally more fruitful. Several private associations of churchmen and theologians promoted understanding between Eastern Orthodoxy and the "Anglo-Catholic" branch of Anglicanism. The Orthodox, however, were reticent in taking any formal step toward reunion before a satisfactory statement on the content of Anglican faith, taken as a whole, could be obtained.

Ecumenical movement

The contemporary ecumenical movement involved the Orthodox Church from the very beginning. Eastern Orthodox representatives took part in the various Life and Work (practical) and Faith and Order (theological) Conferences from the very beginning of this century. One by one the various independent Orthodox Churches joined the World Council of Churches, created in 1948. Often, and especially at the beginning of their participation, Orthodox delegates had recourse to separate statements, which made clear to the Protestant majorities that, in the Orthodox view, Christian unity was attainable only in the full unity of the primitive apostolic faith from which the Orthodox Church had never departed. This attitude of the Orthodox could be understood only if it made sufficiently clear that the truth—which historic Eastern Orthodoxy claims to preserve—is maintained by the Holy Spirit in the church as a whole and not by any individual or any group of individuals on their own right and also that the unity of Christians—which is the goal of the ecumenical movement—does not imply cultural, intellectual, or ritual uniformity but rather a mystical fellowship in the fullness of truth as expressed in eucharistic Communion.

The ecumenical movement, especially since the second Vatican Council, is today much wider than the formal membership of the World Council of Churches. The principle of conciliarism and the readiness of the popes to appear publicly as equals of Eastern patriarchs—as in the meetings between Pope Paul VI and Patriarch Athenagoras in the 1960s—represent significant moves in the direction of a better understanding between Orthodoxy and Roman Catholicism.

The tendency, however, represented by those Western Christians who apparently identify Christianity with various political or social causes has the effect of again widening the traditional gap that has been, in the past, one of the major causes of the break between East and West.

CHURCH, STATE, AND SOCIETY

In the West, after the fall of the Roman Empire, the church assumed the unifying social function that no other individual or institution was able to fulfill. Eventually, the popes were formally invested with civil authority in Christendom. In the East the empire persisted until 1453 and in Russia until 1917; thus the church had to fulfill its social functions in the political framework of the Christian empire.

This historical contrast coincides with a theological polarization: the Eastern Fathers conceived the God–man relationship in terms of personal experience and Communion culminating in deification; Western theology, meanwhile, understood man as autonomous in the secular sphere, although controlled by the authority of the church, which was conceived as vicariously representing God.

Caesaropapism

The Byzantine and Eastern form of church–state relations has often been labelled as caesaropapism, because the hierarchy of the church was, most of the time, deprived of the legal possibility of opposing imperial power. But this label is inaccurate in two aspects: first, it presupposes that the emperor possessed a recognizable power to define the content of the faith, comparable to that of the papacy; and, second, it underestimates the power of the church (as a corporate, transfiguring, and deifying power) that is effective without legal guarantees or statutes. The Byzantine ideal of church–state relations was a "symphony" between the civil and the ecclesiastical functions of Christian society. The abuses of imperial power were frequent, but innumerable examples of popular resistance to those imperial decrees that were considered as detrimental to the faith can be cited. Neither the strong emperors of the 7th century, trying to impose Monophysitism, nor the weakened Palaeologans (13th to 15th century), attempting reunion with Rome, were able to overcome the corporate opposition of Orthodox clergy and laity.

The Byzantine conception of church–state relations was not, however, without major weaknesses. It often led to a de facto identification of the interests of the church with those of the empire. Conceived when both the church and the empire were supranational and, in principle, universal, it gradually evolved into a system that gave a sacred sanction to *national* states. Modern ecclesiastical nationalism, which inhibits relations between Orthodox Churches today, is the outcome of the medieval alliance between the empire and the church.

The *millet* system of the Ottoman Empire

Only after the Turkish occupation of the Balkans was civil authority directly assumed by the Orthodox Church hierarchy in the Middle East. It was granted to it by the new Muslim overlords, who chose to administer their Christian subjects as a separate community, or *millet*, ruled by its own religious leaders. The patriarch of Constantinople was thus appointed by the sultan as head (*millet-bachi*) of the entire Christian population of the Ottoman Empire. Understood by some, especially the Greeks, as the heir of Byzantine emperors and by others, especially the Balkan Slavs and Romanians, as an agent of the hated Turks, the patriarch exercised these powers until the secularization of the Turkish republic by Kemal Atatürk in 1921. By that time, however, he had lost most of his jurisdictional powers because of the establishment of autocephalous churches in Greece, Serbia, Bulgaria, and Romania.

The *millet* system, however, survived in other areas of the Middle East. In Cyprus, for example, the church assumed a leading role in national liberation, and its prestige made Archbishop Makarios the natural leader of the young republic.

The *millet* system and the active political responsibilities that it implied for the church, it should be noted, originated in the Ottoman period only and is not in the spiritual tradition of the Christian East as such. The Russian Church is the most recent example of religious survival without practical social or political involvement.

The Orthodox attitude toward social responsibility in the world constitutes a distinct contribution to the contemporary ecumenical movement. But it will be meaningful only if it is understood in its proper framework—*i.e.*, as an understanding of the Christian faith as a personal spiritual experience of God, which is self-sufficient knowledge of God and which, as such, can lead to an authentically Christian witness in the secularized world.

The practical forms of that witness have varied greatly in history, and Orthodox tradition has placed among the church's saints both hermits and politicians, hesychast monks as well as emperors. According to Serge Bulgakov,

a modern Orthodox theologian, the Orthodox Church accepts "a relativism of means and methods," provided there remains "an absolute and unique goal," which is the Kingdom of God still *to come* but also already *present* in the mystery of the church.

Codification and systematization of practical devices in the fields of personal or social ethics is foreign to Orthodoxy, which rather relies on free human conscience; each Christian, in his behaviour, stands before the judgment of the New Testament and of the great examples of the saints.

(J.M.)

BIBLIOGRAPHY

General: Introductory surveys of the history and doctrines of Eastern Orthodoxy may be found in ERNST BENZ, *The Eastern Orthodox Church, Its Thought and Life* (1963; originally published in German, 1957); TIMOTHY WARE (KALLISTOS WARE), *The Orthodox Church* (1963, reprinted with revisions, 1984); JOHN MEYENDORFF, *The Orthodox Church: Its Past and Its Role in the World Today*, 3rd rev. ed. (1981; originally published in French, 1960), with special attention given to 19th- and 20th-century history; and DEMETRIOS J. CONSTANTELOS, *Understanding the Greek Orthodox Church: Its Faith, History, and Practice* (1982).

History: Overviews of the history of the Eastern Orthodox church are provided by NICOLAS ZERNOV, *Eastern Christendom: A Study of the Origin and Development of the Eastern Orthodox Church* (1961); ALEXANDER SCHMEMANN, *The Historical Road of Eastern Orthodoxy* (1963, reprinted 1977; originally published in Russian, 1954); and JOHN MEYENDORFF, *The Byzantine Legacy in the Orthodox Church* (1982). The best special account of the church in Byzantium until 1261 is found in GEORGE EVERY, *The Byzantine Patriarchate, 451–1204*, 2nd ed. rev. (1962, reprinted 1980). Other works on Byzantine Orthodoxy include STEVEN RUNCIMAN, *The Byzantine Theocracy* (1977), on the relationship between church and state; and J.M. HUSSEY, *The Orthodox Church in the Byzantine Empire* (1986). On the schism between Orthodoxy and Rome, STEVEN RUNCIMAN, *The Eastern Schism* (1955, reprinted 1983); and YVES CONGAR, *After Nine Hundred Years: The Background of the Schism Between the Eastern and Western Churches* (1959, reprinted 1978), are useful. STEVEN RUNCIMAN, *The Great Church in Captivity: A Study of the Patriarchate of Constantinople from the Eve of the Turkish Conquest to the Greek War of Independence* (1968, reissued 1985), critically evaluates the Greek struggle for survival in the Ottoman Empire.

The history of the Orthodox church in Russia is told in VALERI LOBACHEV and VLADIMIR PRAVOTOROV, *A Millennium of Russian Orthodoxy* (1988); GEORGE P. FEDOTOV, *The Russian Religious Mind*, 2 vol. (1946–66, reissued 1975), the best general account of Russian medieval Christianity; DIMITRY POSPIELOVSKY, *The Russian Church Under the Soviet Regime, 1917–1982*, 2 vol. (1984), with a comprehensive bibliography; WILLIAM B. STROYEN, *Communist Russia and the Russian Orthodox Church, 1943–1962* (1967), a sociological analysis; and JANE ELLIS, *The Russian Orthodox Church: A Contemporary History* (1986), from the mid-1960s to the mid-1980s.

Doctrine and sacraments: Doctrinal aspects of the Eastern Orthodox church are detailed by SERGE BULGAKOV, *The Orthodox Church* (1935, reissued 1988; originally published in French, 1932); and VLADIMIR LOSSKY, *The Mystical Theology of the Eastern Church* (1957, reprinted 1976; originally published in French, 1944), a classic on God–man relations, and *Orthodox Theology: An Introduction* (1978). The evolution of Orthodox theology in the Byzantine period is described in JAROSLAV PELIKAN, *The Christian Tradition: A History of the Development of Doctrine*, vol. 2, *The Spirit of Eastern Christendom (600–1700)* (1975); and JOHN MEYENDORFF, *Byzantine Theology: Historical Trends and Doctrinal Themes*, 2nd ed. (1979). GEORGE A. MALONEY, *A History of Orthodox Theology Since 1453* (1976), covers developments since the fall of Constantinople to the Turks in five Orthodox traditions: Russian, Greek, Serbian, Bulgarian, and Romanian. ARTHUR CARL PIEPKORN, *Profiles in Belief: The Religious Bodies of the United States and Canada*, vol. 1, *Roman Catholic, Old Catholic, Eastern Orthodox* (1977), is written for the layperson.

ALEXANDER SCHMEMANN, *For the Life of the World: Sacraments and Orthodoxy*, 2nd rev. and expanded ed. (1973, reissued 1982), is the basic approach to liturgy and sacraments. LEONID OUSPENSKY and VLADIMIR LOSSKY, *The Meaning of Icons*, 2nd ed. (1982; originally published in German, 1952), interprets the icon in its theological and liturgical contexts.

The church and the world: Ancient and modern missions are described by FRANCIS DVORNIK, *Byzantine Missions Among the Slavs: SS. Constantine-Cyril and Methodius* (1970); SERGE BOLSHAKOFF, *The Foreign Missions of the Russian Orthodox Church* (1943); and CONSTANCE J. TARASAR (ed.), *Orthodox America, 1794–1976: Development of the Orthodox Church in America* (1975).

Differences between Orthodoxy and two other Christian denominations are summarized in JOHN MEYENDORFF, *Orthodoxy and Catholicity* (1966; originally published in French, 1965); and CARNEGIE SAMUEL CALIAN, *Icon and Pulpit: The Protestant-Orthodox Encounter* (1968).

Analyses of the relationship between the Eastern Orthodox Church and the state can be found in CHARLES A. FRAZEE, *The Orthodox Church and Independent Greece, 1821–1852* (1969), which uses contemporary diplomatic archives to describe the role of the Greek Church in an event that also involved all major European powers; STEVEN RUNCIMAN, *The Orthodox Churches and the Secular State* (1971), covering the period from the Byzantine Empire to the contemporary situations under Communism, Islām, and dictators; and PEDRO RAMET (ed.), *Eastern Christianity and Politics in the Twentieth Century* (1988).

(Ed.)

Echinoderms

Echinodermata is a phylum of marine invertebrate animals whose living representatives include the classes Crinoidea (sea lilies and feather stars), Echinoidea (sea urchins), Holothuroidea (sea cucumbers), Asteroidea (starfishes, or sea stars), Ophiuroidea (basket stars and serpent stars, or brittle stars), and the recently discovered Concentricycloidea (sea daisies). Beginning with the Lower Cambrian Period almost 570,000,000 years ago, echinoderms have a rich fossil history and are well represented by many bizarre groups, most of which are now extinct.

Echinoderms have been recognized since ancient times; echinoids, for example, were used extensively by Greeks and Romans for medicinal purposes and as food. During the Middle Ages, fossil echinoids and parts of fossil crinoids were objects of superstition. In the early part of the 19th century, Echinodermata was recognized as a distinct group of animals and was occasionally associated with the cnidarians and selected other phyla in a division of the animal kingdom known as the Radiata; the concept of a superphylum called Radiata is no longer valid. Echinoderms now may be separated into 21 classes, based mainly on differences in skeletal structures. The number of extant species exceeds 6,000, and approximately 13,000 fossil species have been described.

For coverage of related topics in the *Macropædia* and *Micropædia*, see the *Propædia*, section 313, and the *Index*.

This article is divided into the following sections:

General features 857
- Size range and diversity of structure
- Distribution and abundance
- Importance

Natural history 858
- Reproduction and life cycle
- Food and feeding habits
- Locomotion
- Ecology

Form and function 861
- General features
- External features
- Internal features

Paleontology and evolution 863
- Extinct echinoderms
- Extant echinoderms

Classification 863
- Distinguishing taxonomic features
- Annotated classification
- Critical appraisal

Bibliography 865

GENERAL FEATURES

Size range and diversity of structure. Although most echinoderms are of small size, ranging up to 10 centimetres (four inches) in length or diameter, some reach relatively large sizes; *e.g.*, some sea cucumbers are as long as two metres (about 6.6 feet), and a few starfishes have a diameter of up to one metre. Among the largest echinoderms were some extinct (fossil) crinoids (sea lilies), whose stems exceeded 20 metres in length.

Echinoderms exhibit a great diversity of body forms, especially among the extinct groups. Although all living echinoderms have a pentamerous (five-part) radial symmetry, an internal skeleton, and a water-vascular system derived from the coelom (central cavity), their general appearance ranges from that of the stemmed, flowerlike sea lilies, to the wormlike, burrowing sea cucumbers, to the heavily armoured intertidal starfish or sea urchin (see Figure 1). The general shape of the echinoderm may be that of a star with arms extended from a central disk or with branched and feathery arms extended from a body often attached to a stalk, or it may be round to cylindrical. Plates of the internal skeleton may articulate with each other (as in sea stars) or be sutured together to form a rigid test (sea urchins). Projections from the skeleton, sometimes resembling spikes, which are typical of echinoderms, give the phylum its name (from Greek *echinos*, "spiny," and *derma*, "skin"). The surface of holothurians, however, is merely warty.

Colour range

Echinoderms also exhibit especially brilliant colours such as reds, oranges, greens, and purples. Many tropical species are dark brown to black, but lighter colours, particularly yellows, are common among species not normally exposed to strong sunlight.

Distribution and abundance. Diverse echinoderm faunas consisting of many individuals and many species are found in all marine waters of the world except the Arctic, where few species occur. Echinoids, including globular spiny urchins and flattened sand dollars, and asteroids are commonly found along the seashore. Although many species are restricted to specific temperate regions, Arctic, Antarctic, and tropical forms often are widely distributed; many species associated with coral reefs, for example, range across the entire Indian and Pacific oceans. Many of the echinoderms of Antarctica are distributed around the continent; those with a floating (planktonic) larval stage may be widely distributed, carried great distances by ocean currents. Some species, particularly those in Antarctic and deep-sea regions, have achieved a wide distribution without benefit of a floating larval stage. They may have done so by migration of adults across the seafloor or, in the case of shallow-water species, by passive transport across oceans in rafts of seaweed. Echinoderms tend to have a fairly limited depth range; species occurring in near-shore environments do not normally reach depths greater than 100 metres. Some deep-sea species may be found over a considerable range of depths, often from 1,000 metres to more than 5,000 metres. One sea cucumber species has a known range of 37–5,205 metres. Only sea cucumbers reach ocean depths of 10,000 metres and more.

Depth range

Importance. *Role in nature.* Echinoderms are efficient scavengers of decaying matter on the seafloor, and they prey upon a variety of small organisms, thereby helping to regulate their numbers. When present in large numbers, sea urchins can devastate sea-grass beds in the tropics, adversely affecting the organisms dwelling within. Sea urchins that burrow into rocks and along a shore can accelerate the erosion of shorelines. Other tropical species of sea urchins, however, control the growth of seaweeds in coral reefs, thereby permitting the corals to flourish. Removal of the sea urchins results in the overgrowth of seaweeds and the devastation of the coral reef habitat. Echinoderms can alter the structure of seafloor sediments in a variety of ways. Many sea cucumbers feed by swallowing large quantities of sediment, extracting organic matter as the sediment passes through the intestine, and ejecting the remainder. Large populations of sea cucumbers in an area can turn over vast quantities of surface sediments and can greatly alter the physical and chemical composition of the sediments. Burrowing starfish, sand dollars, and heart urchins disturb surface and subsurface sediments, sometimes to depths of 30 centimetres or more. In addition, echinoderms produce vast numbers of larvae that provide food for other planktonic organisms.

Relation to human life. Some of the larger species of

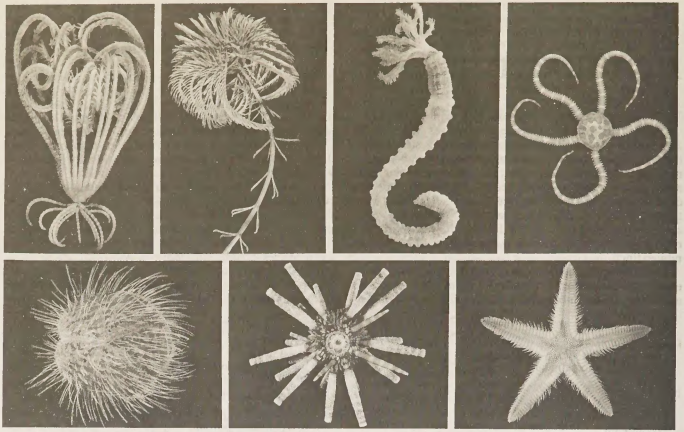


Figure 1: *Representative extant echinoderms.*
(Top left) Feather star (*Crinometra brevipinna*). Slender cirri at the lower end of the animal are used for attachment to a substrate. (Top centre left) Stalked crinoid (*Neocrinus decorus*). The stalk attaches to a hard substrate. (Top centre right) Holothurian (*Euapta lappa*), which has a soft body, with 10 or more feeding tentacles around the mouth. (Top right) Brittle star (*Ophiomyxa flaccida*) with long slender arms. (Bottom left) An irregular echinoid (*Lnopneustes longispinus*). (Bottom centre) A regular echinoid (*Eucidaris tribuloides*) with spines and test with five-part symmetry. (Bottom right) Asteroid (*Astropecten nitidus*), showing typical five-part symmetry.
John E. Miller

tropical sea cucumbers, known commercially as trepang or bêche-de-mer, are dried and used in soups, particularly in Asia. Raw or cooked mature sex organs, or gonads, of sea urchins are regarded as a delicacy in some parts of the world, including parts of Europe, the Mediterranean region, Japan, and Chile. Some tropical holothurians produce a toxin, known as holothurin, which is lethal to many kinds of animals; Pacific islanders kill fish by poisoning waters with holothurian body tissues that release the toxin. Holothurin does not appear to harm human beings; in fact, the toxin has been found to reduce the rate of growth of certain types of tumours and thus may have medical significance. The eggs and spermatozoa of echinoderms, particularly those of sea urchins and starfishes, are easily obtained and have been used to conduct research in developmental biology. Indeed, echinoids have been collected in such large numbers that they have become rare or have disappeared altogether from the vicinity of several marine biologic laboratories.

Holothurin toxin

Starfishes that prey upon commercially usable mollusks, such as oysters, have caused extensive destruction of oyster beds. Sea urchins along the California coast have interfered with the regrowth of commercial species of seaweed by eating the young plants before they could become firmly established. The crown-of-thorns starfish, which feeds on living polyps of reef corals, has caused extensive short-term damage to coral reefs in some parts of the Pacific and Indian oceans.

NATURAL HISTORY

Reproduction and life cycle. In most species the sexes are separate; *i.e.*, there are males and females. Although reproduction is usually sexual, involving fertilization of eggs by spermatozoa, several species of sea cucumbers, starfishes, and brittle stars can also reproduce asexually.

Asexual reproduction. Asexual reproduction in echinoderms usually involves the division of the body into two or more parts (fission) and the regeneration of missing body parts. Fission is a common method of reproduction in approximately 60 species of asteroids, ophiuroids, and holothurians (a total of less than 1 percent of the living species), and in some of these species sexual reproduction is not known to occur. Successful fission and regeneration require a body wall that can be torn and an ability to seal resultant wounds. In some asteroids fission occurs when two groups of arms pull in opposite directions, thereby tearing the animal into two pieces. Successful regeneration requires that certain body parts be present in the lost pieces; for example, many asteroids and ophiuroids can regenerate a lost portion only if some part of the disk is present. In sea cucumbers, which divide transversely, considerable reorganization of tissues occurs in both regenerating parts.

Fission and regeneration

The ability to regenerate, or regrow, lost or destroyed parts is well developed in echinoderms, especially sea lilies, starfishes, and brittle stars, all of which can regenerate new arms if existing ones are broken off. Echinoderm regeneration frustrated early attempts to keep starfishes from destroying oyster beds; when captured starfishes were chopped into pieces and thrown back into the sea, they actually increased in numbers. So long as a portion of a body, or disk, remained associated with an arm, new starfishes regenerated. Some sea cucumbers can expel their internal organs (autoevisceratе) under certain conditions (*i.e.*, if attacked, if the environment is unfavourable, or on a seasonal basis), and a new set of internal organs regenerates within several weeks. Sea urchins (Echinoidea) readily regenerate lost spines, pincerlike organs called pedicellariae, and small areas of the internal skeleton, or test.

Sexual reproduction. In sexual reproduction, eggs (up to several million) from females and spermatozoa from males are shed into the water (spawning), where the eggs are fertilized. Most echinoderms spawn on an annual cycle, with the spawning period normally lasting one or two months during spring or summer; several species, however, are capable of spawning throughout the year. Spawn-inducing factors are complex and may include external influences such as temperature, light, or salinity of the

water. In the case of one Japanese feather star (Crinoidea), spawning is correlated with phases of the Moon and takes place during early October when the Moon is in the first or last quarter. Many echinoderms aggregate before spawning, thus increasing the probability of fertilization of eggs. Some also display a characteristic behaviour during the spawning process; some asteroids and ophiuroids raise the centre of the body off the seafloor; holothurians may raise the front end of the body and wave it about. These movements are presumably intended to prevent eggs and sperm from becoming entrapped in the sediment.

Development. After an egg is fertilized, the development of the resulting embryo into a juvenile echinoderm may proceed in a variety of ways. Small eggs without much yolk develop into free-swimming larvae that become part of the plankton, actively feeding on small organisms until they transform, or metamorphose, into juvenile echinoderms and begin life on the seafloor. Larger eggs with greater amounts of yolk may develop into a larval form that is planktonic but subsists upon its own yolk material, rather than feeding upon small organisms, before eventually transforming into a juvenile echinoderm. Development involving an egg, planktonic larval stages, and a juvenile form is termed indirect development. Echinoderm development in which large eggs with abundant yolk transform into juvenile echinoderms without passing through a larval stage is termed direct development.

Brooding

In direct development the young usually are reared by the female parent. Parental care or brood protection ranges from actual retention of young inside the body of the female until they are born as juveniles to retention of the young on the outer surface of the body. Brood protection is best developed among Antarctic, Arctic, and deep-sea echinoderms, in which young may be held around the mouth or on the underside of the parent's body, as in some starfishes and sea cucumbers, or in special pouches on the upper surface of the body, as in some sea urchins, sea cucumbers, and asteroids.

During indirect development, the fertilized egg divides many times to produce a hollow ciliated ball of cells (blastula); cleavage is total, indeterminate, and radical. The blastula invaginates at one end to form a primitive gut, and the cells continue to divide to form a double-layered embryo called the gastrula. Echinoderms resemble vertebrates and some invertebrate groups (chaetognaths and hemichordates) in being deuterostomes; the hole through which the gut opens to the outside (blastopore) marks the position of the future anus; the mouth arises anew at the opposite end of the body from the blastopore. A pair of subdivided hollow pouches arise from the gut and develop into the body cavity (coelom) and water-vascular system.

Larval types

The gastrula develops into a basic larval type called a dipleurula larva, characterized by bilateral symmetry; hence the name, which means "little two sides." A single band of hairlike projections, or cilia, is found on each side of the body and in front of the mouth and anus. The characteristic larvae found among the living classes of echinoderms (Figure 2) are modifications of the basic dipleurula pattern.

Because the ciliated band of the dipleurula larva of holothurians becomes sinuous and lobed, thus resembling a human ear, the larva is known as an auricularia larva. The dipleurula larva of asteroids develops into a bipinnaria larva with two ciliated bands, which also may become sinuous and form lobes or arms; one band lies in front of the mouth, the other behind it and around the edge of the body. In most asteroids the larval form in the next stage of development is called a brachiolaria, which has three additional arms used for attaching the larva to the seafloor. Echinoids and ophiuroids have complex advanced larvae closely similar in type. The larva, named pluteus, resembles an artist's easel turned upside down. It has fragile arms formed by lobes of ciliated bands and is supported by fragile rods of calcite, the skeletal material. The echinoid larva (echinopluteus) and the ophiuroid larva (ophiopluteus) usually have four pairs of arms but may have fewer or more. An extra unpaired arm on the plutei of sand dollars and cake urchins extends downward, presumably to help keep the larva upright. The crinoids, which apparently lack a dipleurula larval stage, have a barrel-shaped larva called a doliolaria larva. The doliolaria larva also occurs in other groups; in holothurians, for example, it is the developmental stage after the auricularia larva, which may not occur in some species. A doliolaria larva usually contains large quantities of yolk material and moves with the aid of several ciliated bands arranged in hoops around the body.

Although most larval stages are small, often less than one millimetre (0.04 inch) in length, some holothurians are known to be 15 millimetres long in the larval stage, and the length of bipinnaria larvae of some starfishes may exceed 25 millimetres.

After a few days to several weeks in a free-swimming form (plankton), echinoderm larvae undergo a complex transformation, or metamorphosis, that results in the juvenile echinoderm. During metamorphosis, the fundamental bilateral symmetry is overshadowed by a radial symmetry dominated by formation of five water-vascular canals (see below *Form and function: External features*). Among holothurians, echinoids, and ophiuroids, the larvae may metamorphose as they float, and the young then sink to the seafloor; among crinoids and asteroids, however, the larvae firmly attach to the seafloor prior to metamorphosis. The average life span of echinoderms is about four years, and some species may live as long as eight or 10.

Metamorphosis

Food and feeding habits. Echinoderms feed in a variety of ways. A distinct feeding rhythm frequently occurs, with many forms feeding only at night, others feeding continuously. Feeding habits range from active, selective predation to omnivorous scavenging or nonselective mud swallowing.

Crinoids are suspension feeders, capturing planktonic organisms in a network of mucus produced by soft appendages, called tube feet, contained in grooves on the tentacles, or arms. The arms are spread into a characteristic "fan" at right angles to the prevailing current, and small prey animals are passed to the mouth along the grooves by activity of the cilia and the tube feet.

Many asteroids are active predators on shellfishes and even upon other starfishes; other asteroids are mud swal-

From J.A. Pechenik, *Biology of the Invertebrates* (1985); reproduced by permission of PWS Publishers

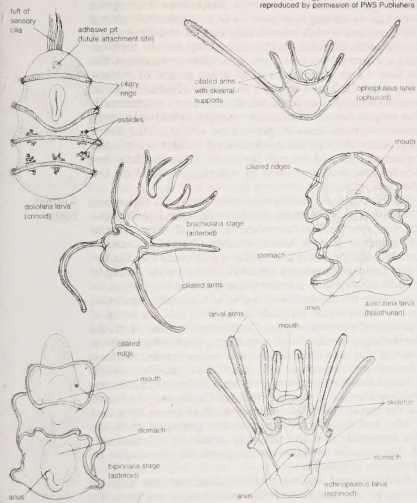


Figure 2: Larval forms of some living classes of echinoderms.

lowers. When feeding, some asteroid species extrude their stomach through the mouth onto the prey, which then is partially digested externally, after which the stomach is retracted and digestion is completed inside the body. Most ophiuroids feed on small organisms floating in the water or lying on the bottom, which are captured by the arms and tube feet and passed toward the mouth. Ophiuroids with arms branched in a complex manner may feed in a way similar to that of the crinoids. Feeding methods of concentricycloids are not yet known.

Regular and irregular echinoids

The more primitive, so-called regular, sea urchins are omnivorous or vegetarian browsers, either scraping algae and other small organisms from rocks with their hard teeth or eating seaweed. Several deep-sea regular echinoids feed exclusively on plants carried into the sea from the land. The more advanced irregular echinoids, which usually lack teeth, are burrowers and pass small organisms to the mouth with the aid of spines and tube feet. Several species of sand dollars sometimes feed on suspended organisms carried to them by ocean currents as they lie on the seafloor. Some sea cucumbers remain attached to a surface for indefinite periods of time, capturing plankton in a network of branching, sticky tentacles; others select food from the seafloor and push it into their mouths with their tentacles. A large number of holothurians feed by actively swallowing mud and sand, digesting the organic material, and egesting the waste in the form of characteristic castings, in a manner similar to that of earthworms.

Under artificial conditions, as in aquariums, echinoderms can survive apparent starvation for several weeks at a time. After a holothurian has autoeviscerated, it is unable to feed during the several weeks required for gut regeneration. Echinoderms may derive a significant amount of nourishment, at least for the outer cell layers of the body, from organic material dissolved in seawater.

Locomotion. Asteroids and echinoids, which use spines and tube feet in locomotion, may move forward with any area of the body and reverse direction without turning around. The feet may be used either as levers, by means of which the echinoderm steps along a surface, or as attachment mechanisms that pull the animal. Sea daisies presumably move in the same way. Ophiuroids tend to move by thrashing the arms in one of several possible methods, including a rowing motion in which strokes are taken by two pairs of extended arms; the fifth arm either is extended forward in the direction in which the animal is traveling or trails behind.

Holothurians (sea cucumbers) generally lead with the mouth, or oral, end, movement being carried out by both the tube feet and contraction and expansion of the body; sluglike movement is common. Holothurians of the family Synaptidae are able to pull themselves across a surface using their sticky tentacles as anchors.

Stalked crinoids (sea lilies), so called because they have stems, generally are firmly fixed to a surface by structures at the ends of the stalks called holdfasts. Some fossil and living forms release themselves to move to new attachment areas. The unstalked crinoids (feather stars) generally swim by thrashing their numerous arms up and down in a coordinated way; for example, in a 10-armed species, when arms 1, 3, 5, 7, and 9 are raised upward, arms 2, 4, 6, 8, and 10 are forcibly pushed downward; then the former group of arms thrashes downward as the latter is raised. Feather stars that do not swim pull themselves across a surface using their arms.

Swimming

Swimming is known to occur in crinoids, ophiuroids, and holothurians. Some holothurians, formerly regarded as strictly bottom-living forms, are capable of efficient swimming; others, with gelatinous or flattened bodies and reduced calcareous skeletons, spend most of their lives swimming in deep water.

Righting response. Among echinoderms a normal position may be with the mouth either facing a surface, as in asteroids, ophiuroids, concentricycloids, and echinoids, or facing away from it, as in crinoids and holothurians. When overturned, echinoderms exhibit a righting response. Starfishes show this response most effectively, using the tube feet and the arms to perform a slow, graceful somersault that restores their normal position. Sea urchins roll themselves over by a concerted action of their tube feet and spines. The flat sand dollar can turn itself over only by burrowing into the sand until its position is vertical, then toppling over. In more agile groups such as holothurians, crinoids, and ophiuroids, righting is performed with relative ease.

Burrowing. Many echinoderms burrow in rock or soft sediments. Crinoids do not burrow because their feeding apparatus must be kept clear of sediment. Some urchins use the combined abrasive actions of their spines and teeth to burrow several inches into rock, usually in areas of severe wave and tidal action. The so-called irregular echinoids excavate soft sediments to various depths; most sand dollars burrow just below the surface, and some heart urchins may be found at depths of 38 centimetres or more. Holothurians use tentacles and contraction of the body wall in burrowing that generally is related to feeding. Several asteroid species bury themselves in sandy or muddy areas. The characteristic position of several ophiuroid groups involves burying the body into a surface and leaving only the tips of the arms projecting for food gathering.

Ecology. *Habitats.* Echinoderms are exclusively marine animals, with only a few species tolerating even brackish water. Among the exceptions are a few tropical holothurians that can withstand partial drying if stranded on a beach by a receding tide. Most echinoderms cannot tolerate marked changes in salinity, temperature, and light intensity and tend to move away from areas where these factors are not optimal. The behaviour of a large proportion of shallow-water species is regulated by light; *i.e.,* individuals remain concealed during the day and emerge from concealment at night for active feeding. Echinoderms are found in the warmest and coldest of the world's seas; those species that can tolerate a broad temperature range usually also have a broad geographic range. The horizontal or vertical distribution of many species is also governed by water temperature. The influence of pressure upon echinoderms has not yet been thoroughly investigated.

Influence of external factors

Echinoderms occupy a variety of habitats. Along a rocky shore, starfishes and sea urchins may cling to rocks beneath which sea cucumbers and brittle stars are concealed. Some sea urchins have special adaptations for coping with surf pounding against rocks (*e.g.,* particularly strong skeletons and well-developed tube feet for attachment). In sandy areas starfishes, brittle stars, irregular sea urchins, and sea cucumbers may bury themselves or move on the surface. Large populations of all living groups of echinoderms can be found in mud and ooze offshore. In some marine areas, echinoderms are the dominant organism; in the deepest ocean trenches, for example, holothurians may constitute more than 90 percent by weight of the living organisms. Perhaps the most unusual habitat is exploited by sea daisies and a small family of asteroids; these animals occur only on pieces of waterlogged wood on the deep-sea floor.

Animal hosts

Echinoderms frequently use other animals as homes; thousands of brittle stars, for example, may live in some tropical sponges. Sea cucumbers may attach themselves to the spines of sluggish Antarctic echinoids, and one sea cucumber attaches itself to the skin of a deep-sea fish. On the other hand, echinoderms are also hosts to a wide variety of organisms. Various crustaceans and barnacles, for example, cause the formation of galls, or tumourlike growths, in the skeletons of sea urchins, and crinoids are hosts of specialized parasitic worms. Commensal worms, which do no damage, are associated with most groups; an interesting case of commensalism is the association between various tropical sea cucumbers and the slender pearlfish, which often is found in the rectum of the holothurian, head protruding through its anus. Pinnotherid crabs may be found in the rectum of echinoids and holothurians in Peru and Chile, and highly modified parasitic gastropod mollusks are frequently found in the body cavities of holothurians. A conspicuous parasitic sponge grows on two species of Antarctic ophiuroids.

Predation and defense. Although echinoderm populations do not generally suffer from heavy predation by other animals, ophiuroids form a significant part of the

diet of various fishes and some asteroids. Echinoids are frequently eaten by sharks, bony fishes, spider crabs, and gastropod mollusks; crows, herring gulls, and eider ducks may either peck their tests (internal skeletons) or drop them repeatedly until they break; and mammals, including the Arctic fox, sea otters, and humans, eat them in considerable numbers. Asteroids are eaten by other asteroids, mollusks, and crustaceans. Some holothurians are eaten by fishes and by humans. Crinoids appear to have no consistent predators.

Defense mechanisms

Echinoderms can protect themselves from predation in a variety of ways, most of which are passive. The presence of a firm skeleton often deters predators; echinoids, for example, have a formidable array of spines and, in some cases, highly poisonous stinging pincerlike organs (pedicellariae), some of which may cause intense pain and fever in humans. Some asteroids use chemical secretions to stimulate violent escape responses in other animals, particularly predatory mollusks. Some holothurians eject from the anus a sticky mass of white threads, known as cuvierian tubules, which may entangle or distract predators; others produce holothurin, a toxin lethal to many would-be predators.

Aggregation. Echinoderms tend to aggregate in large numbers and evidently also did so in the past; fossil beds consisting almost exclusively of large numbers of one or a few species are known from as early as the Lower Cambrian. In present-day seas, ophiuroids may cover large areas of the seafloor; vast aggregations of echinoids are also common. Holothurians, crinoids, and some asteroids also often show a tendency to aggregate.

The phenomenon of aggregation apparently is a response to one or more environmental factors, chief of which is availability of food; *e.g.*, large numbers of ophiuroids and crinoids occupy areas in which strong currents carry large amounts of plankton. An ophiuroid raises some arms into the water to capture food, using other arms to hold on to other nearby ophiuroids; in this way, a large aggregation can maintain its position in an environment in which a single ophiuroid or a small clump of them would be swept away. As stated previously, aggregation also enhances possibilities for successful propagation of a species and possibly may afford some protection from predators. Aggregation may be a passive phenomenon resulting from interactions between individuals and the environment as well as a demonstration of true social behaviour, a result of interactions among individuals.

FORM AND FUNCTION

Coelom

General features. Echinoderms have a skeleton composed of numerous plates of mineral calcium carbonate (calcite). Part of the body cavity, or coelom, is a water-vascular system, consisting of fluid-filled vessels that are pushed out from the body surface as tube feet, papillae, and other structures that are used in locomotion, feeding, respiration, and sensory perception. The conspicuous five-rayed, or pentamerous, radial symmetry of living echinoderms tends to obliterate their fundamental bilateral symmetry.

External features. *Symmetry and body form.* Many of the earliest echinoderms either lacked symmetry or were bilaterally symmetrical. Bilateral symmetry occurs in all living groups and is especially marked in the larval stages. A tendency toward radial symmetry (the arrangement of body parts as rays) developed early in echinoderm evolution and eventually became superimposed upon the fundamental bilateral symmetry, often obliterating it. Radial pentamerous symmetry is conspicuous among all groups of living echinoderms. Although the reasons for the success of radial symmetry are not yet completely understood, it has been suggested that a pentamerous arrangement of skeletal parts strengthens an animal's skeleton more than would, for example, a three-rayed symmetry.

Pentamerous structure is evident in the arrangement of the tube feet, which usually radiate from the mouth in five bands. Many of the major organ systems, including the water-vascular system, muscles, hemal system (a series of fluid-filled spaces of indeterminate function), and parts of the nervous system are also pentamerous. The skeleton follows a pentamerous pattern, except in holothurians, where it is usually reduced to microscopic ossicles (bones).

Growth patterns

Distinct growth patterns among the echinoderms provide some basis for separating the phylum into subphyla. Among homalozoans, the pattern is asymmetrical. In crinozoans and blastozoans, bands of tube feet radiate from the mouth, cross the theca (*i.e.*, sheath or calyx), and extend onto the brachioles or arms; in asterozoans the bands of tube feet radiate outward from the mouth onto the arms to produce a star shape; and in the armless echinozoans, the tube feet form five meridians on the spherical or cylindrical body.

The crinoid (sea lily, feather star) mouth is centrally located on the cup-shaped theca; from which arise a variable number of arms resembling fern fronds. Although five is the primitive number of arms, they branch once or several times in most living forms to produce 10 to 200 arms. Crinoids either are supported on a stem (or stalk) attached to the underside of the theca, or they lack stalks, as is the case with most living forms, and attach themselves by means of slender appendages adapted for grasping (prehensile cirri).

Asteroids have a large central disk from which radiate five or more hollow arms containing parts of the major internal organ systems. The underside (oral surface) of the disk contains a centrally located mouth; the underside of each arm contains five or more bands of tube feet in special grooves called ambulacral furrows. The upper (aboral) surface of the disk has a centrally located anus (often absent) and the sieve plate (madreporite) of the water-vascular system (see below *Internal features: Water-vascular system*). Seven-armed starfish species are not unusual, a deep-sea family has six to 20 arms, and one Antarctic genus may have up to 50 arms. Concentricycloids have a discoid body; the dorsal surface is plated andthe ventral surface is naked (Figure 3).

Ophiuroids have a small disk from which five arms radiate. The larger internal organs usually are confined to the disk. The centrally located mouth is on the underside of the disk as are the tube feet, which are not arranged in special grooves. Although most ophiuroids have five arms, a few have six or more, and in one group, the basket stars, the five arms are branched to form a complex network.

In echinoids (Figure 4) the skeleton forms a rounded,

From A.N. Baker, F.W.E. Rowe, and H.E.S. Clark, *Nature* (26 June 1986); reprinted by permission from *Nature*, vol. 321, no. 6073, pp. 862–864. copyright © 1986 Macmillan Journals Limited

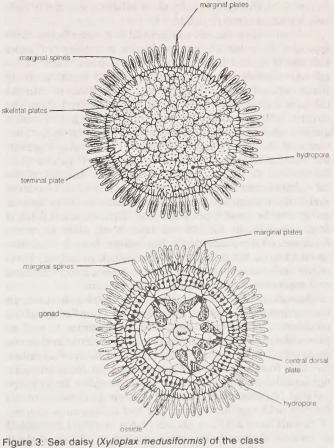


Figure 3: Sea daisy (*Xyloplax medusiformis*) of the class Concentricycloidea. (Top) Dorsal view; (bottom) ventral view.

or globular, test of solid plates; tube feet, which emerge through holes in the plates, form five conspicuous bands, or ambulacra. Spaces between bands of feet are called interambulacra. Regular echinoids are roughly spherical in shape, with a centrally located mouth at the junction of the five bands containing tube feet (ambulacra); the anus is located on the side of the body opposite the mouth (aboral). Irregular urchins are elongated or flattened in shape, with the anus on the oral or aboral surface of the body. In regular and some irregular echinoids, the mouth is equipped with five teeth operated by a complex system of plates and muscles called Aristotle's lantern.

Aristotle's lantern

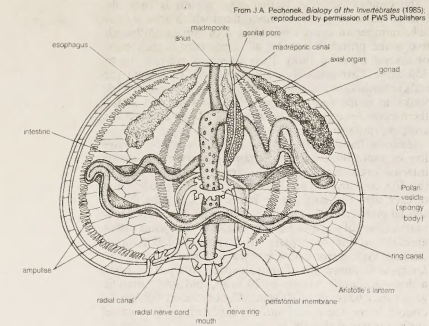


Figure 4: Sea urchin (*Arbacia punctulata*). Internal structures are shown laterally in diagrammatic section.

Holothurians are elongated, with mouth and anus at opposite ends of the body. The spaces between the tube feet, which are arranged in five rows, or radii, are known as interradii. The tube feet may be more numerous on the underside of the body than elsewhere, scattered over radii and interradii, or absent. Most holothurians are soft-bodied animals because the skeleton is reduced and the skeletal units, called ossicles or spicules, are microscopic in size. Holothurians usually show bilateral symmetry outside, radial symmetry inside.

Skeleton. The skeleton is dermal but nonetheless conspicuous in echinoderms, with the exception of most holothurians, and forms an effective armour. Each skeletal unit (ossicle) usually consists of two parts, a living tissue (stroma) and a complex lattice (stereom) of mineral calcium carbonate, or calcite, which is derived from the stroma. In living echinoderms, certain properties of calcite are not evident in the stereom because of its latticed structure and the presence of soft stroma. In fossils, however, the stroma may be replaced by secondary calcite (*i.e.*, calcite laid down in continuity with the original skeletal calcite), and recognition of fragments of echinoderm skeletons in fossil strata is easier because no other animal group has the same type of skeleton. Each ossicle is formed from granules in the dermal layer that, after secretion from special lime-secreting cells, enlarge, branch, and fuse to build up a three-dimensional network of calcite. Parts of the skeleton enlarge as an animal grows, and resorption and regeneration of the skeleton may occur.

Skeletal structures

Echinoderms exhibit a variety of skeletal structures. In the echinoids, a hollow test (skeleton) consisting of 10 columns of plates bears large and small spines as well as pincerlike organs (pedicellariae) used in defense and in the removal of unwanted particles from the body. Pedicellariae, also found in the asteroids, are absent from crinoids, ophiuroids, and holothurians. The complex feeding apparatus (Aristotle's lantern) of echinoids consists of 40 ossicles held together by muscles and collagenous sutures.

Crinoids have a hollow sheath (theca or calyx) composed of two or three whorls, each consisting of five skeletal plates; the stalk and the slender appendages (cirri) of unstalked forms consist of a series of drum-shaped ossicles. The asteroid skeleton is composed of numerous smooth or spine-bearing ossicles of various shapes held together by muscles and ligaments, permitting flexibility. The arms of asteroids are hollow, those of ophiuroids solid, with the central axis of each arm consisting of elongated ossicles called vertebrae. The microscopically sized ossicles of holothurians are highly variable in form, ranging from flat lattice plates with holes to exquisitely symmetrical wheels, and are usually numerous; one tropical species, for example, has more than 26,000,000 ossicles in its body wall. A ring of plates, called the calcareous ring, surrounds the tube leading from the mouth to the stomach (*i.e.*, the esophagus) of holothurians. Although located in a similar position to that of the echinoid Aristotle's lantern, the calcareous ring functions as a point of insertion for muscles, not as a feeding apparatus.

Internal features. *Water-vascular system.* The water-vascular system, which functions in the movement of tube feet, is a characteristic feature of echinoderms, and evidence of its existence has been found in even the oldest fossil forms (Figure 5). It comprises an internal hydraulic system of canals and reservoirs containing a watery fluid, the system consisting of a sieve plate, or madreporite, and a ring vessel, or water-vascular ring, that are connected by a frequently calcified vessel called the stone canal. Five radial water canals extend outward from the ring vessel and give rise to branches that end in the tube feet, which are in contact with the sea. The ring vessel in ophiuroids, asteroids, concentricycloids, and holothurians has bulbous cavities called Polian vesicles, which apparently maintain pressure in the system and hold reserve supplies of fluid; ophiuroids have four or more vesicles, asteroids five, holothurians from one to 50. Crinoids lack Polian vesicles, and echinoids have five structures known as either Polian vesicles or spongy bodies.

Polian vesicles

The madreporite, which is usually located externally, takes in water from outside the body; if internally located, as is the case in many holothurians, fluid is taken from the body cavity. The water or fluid passes from the madreporite to the ring vessel and along the radial canals to the tube feet. The tube feet are extended by contractions of localized muscle areas in the radial canals (ophiuroids) or by contractions of offshoots of the radial canals called ampullae (asteroids, concentricycloids, echinoids, and holothurians); the contractions force fluid into the tube feet, which then extend.

The structure of the system varies from group to group; asteroids frequently have more than one madreporite, and in holothurians, the madreporite is usually internal, hanging in the coelom. Radial canals may lie inward or outward from the skeleton. The tube feet may have well-developed suckers with great holding power, may taper to a point, or may be adapted for respiration, feeding, burrow building, mucus production, or sensory perception. Attachment of tube feet to hard substrates is achieved through a combination of suction and mucus production. The mucus contains adhesive and de-adhesive mucopolysaccharides. Respiratory tube feet have high oxygen uptake; they are usually located on parts of the body where

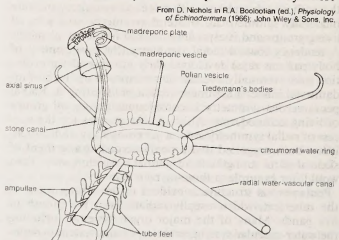


Figure 5: Water-vascular system of an asteroid.

water flow is unimpeded. Tube feet have been implicated in photoreception and chemoreception; the eyespots in the terminal tentacles of asteroids are the most conspicuous photoreceptors.

The tube feet of crinoids are arranged in clumps of three on the arms and on the pinnules. They secrete and spread a net of sticky mucus that traps small organisms. In ophiuroids the tube feet are used to gain a hold on a surface and to pass food to the mouth. The numerous tube feet of asteroids are used in locomotion; asteroids with suckered feet may use them to exert a continuous pull on the valves of shellfish (*e.g.*, oysters, mussels) until muscles holding the valves tire and open slightly, allowing the asteroid to insert its stomach. In sea daisies the ring of tube feet is probably used for attachment to substrates. Holothurians use tube feet for the same purpose. Tentacles around the mouth of holothurians are modified tube feet used to capture food; tentacles used to capture plankton are branched and sticky, while those used to scoop mud and shovel it into the mouth have a simpler structure.

The tube feet of echinoids serve a variety of functions. The mouth of regular echinoids is surrounded by sensory tube feet, and tube feet farther from the mouth are used in locomotion. On the upper side of the body near the anus, the tube feet have respiratory and sensory functions. The tube feet of irregular echinoids, which burrow, are modified in various ways for feeding, burrow construction, and sensory and respiratory functions.

Cilia

Body wall and body cavity. The outer body wall (epidermis) contains hairlike projections (cilia) in most echinoderms except ophiuroids; the body wall of crinoids has relatively few. The cilia produce a waving motion that carries food particles toward the mouth or removes unwanted particles from the body. The epidermis also contains glandular and sensory cells. The epidermis of skeletal elements such as spines and pedicellariae, which project from the body surface, often is worn away. The next layer, the dermis, includes the calcareous skeleton and connective tissues. Internal to the dermis are circular and longitudinal muscle layers. The extensive body cavity (coelom) is modified to form several specialized regions. Two subdivisions of the coelom are the perivisceral coelom and the water-vascular system. The perivisceral coelom is a large, fluid-filled cavity in which the major organs, particularly the digestive tube and sex organs, are suspended. Other regions of the coelom include the axial sinus (absent from adult holothurians and all echinoids), the madreporic vesicle, and the hyponeural sinus (often called the perihemal system).

Alimentary and blood systems. The digestive canal consists of a tube, which is almost straight (asteroids and ophiuroids), coiled in a clockwise direction (crinoids and holothurians), or coiled first clockwise, then counterclockwise (echinoids). The tube may be divided into esophagus, stomach, intestine, and rectum. Specialized branches of the digestive tube enlarge the digestive surface and may serve other functions; *e.g.*, digestive glands of asteroids, diverticula of echinoids and crinoids, siphons in echinoids, and respiratory trees in holothurians. The anus, absent in ophiuroids and a few asteroids, is present in most groups. The mouth is near the centre of the oral surface, at the point of convergence of the areas containing the tube feet.

Hemal ring

The blood system is a complex system of spaces that are neither part of the coelom nor true vessels. A hemal ring and five radial hemal canals surround the esophagus and radial canals of the water-vascular system. A sixth hemal space arises from the hemal ring and enters the axial organ. In addition, a complex network of hemal spaces is associated with the alimentary canal and gonads.

Axial organ. The axial organ, a complex and elongated mass of tissue found in all echinoderms except holothurians, represents the common junction of the perivisceral coelom, the water-vascular system, and the hemal system. Although its functions are not yet well understood, the axial organ plays a part in defense against invading organisms, can contract, is responsible for a circulation of fluids, and may have excretory and secretory activity.

Nervous system and sense organs. The echinoderm nervous system is complex. In all groups, a nerve plexus lies within and below the skin. In addition, the esophagus is surrounded by one to several nerve rings, from which run radial nerves often in parallel with branches of the water-vascular system. Ring and radial nerves coordinate righting activity.

Although echinoderms have few well-defined sense organs, they are sensitive to touch and to changes in light intensity, temperature, orientation, and the surrounding water. The tube feet, spines, pedicellariae, and skin respond to touch, and light-sensitive organs have been found in echinoids, holothurians, and asteroids.

Number of gonads

Reproductive system. The masses of sex cells that compose the gonads of crinoids fill special cavities in the arms or pinnules. Crinoids are the only echinoderms with gonads outside the main body cavity, probably because its volume is reduced. Asteroids typically have 10 gonads, two in each arm, which are located near the arm base, appearing as a feathery tuft or a mass of tubules resembling a bunch of grapes. The gonads of some species are arranged in rows along each arm. Concentricycloids have five pairs of saclike gonads (Figure 3). Ophiuroids have gonads attached to sacs that hang into the body cavity; the sacs, which open outside the body at the bases of the arms, may have one gonad or as many as 1,000.

Regular echinoids typically have five gonads attached to the interambulacra. A duct from each gonad opens to the exterior near the anus. Most irregular echinoids have four gonads, some have three or five, a few have two; the ducts are on the upper surface of the body. Holothurians differ from all other living echinoderms in having a single gonad, which consists of branching or unbranched tubules; the tubules open into a single duct, which opens to the exterior near the front end of the body. Since many early fossil echinoderms have a single genital opening, or gonopore, it is assumed that these forms also had only one gonad; the condition in holothurians thus is regarded as primitive.

PALEONTOLOGY AND EVOLUTION

Extinct echinoderms. Because the phylum Echinodermata was already well diversified by the Lower Cambrian Period, a considerable amount of Precambrian evolution must have taken place. A Precambrian fossil from Australia has triradiate symmetry and a superficial resemblance to an edrioasteroid; it has been suggested that the triradiate condition may have been a precursor of pentamerous symmetry, and that this fossil is a "pre-echinoderm." Scientists speculate that the lack of Precambrian fossil echinoderms indicates that while the earliest echinoderms may have possessed a water-vascular system, they lacked a calcite skeleton and thus did not fossilize. While the fossil record of echinoderms is extensive, there are many gaps, and many questions remain concerning the early evolution of the group. Ancient echinoderms exhibited an extraordinary variety of bizarre body forms; the earliest classes seemed to be "experimenting" with body shapes and feeding mechanisms; most were relatively short-lived.

Adaptation

Early echinoderms were adapted to life on the surface of hard or soft seafloors, though the burrowing habit may have been acquired relatively early by sea cucumbers.

Extant echinoderms. Relationships among the living classes of echinoderms have been the subject of debate for many decades. Some scientists believe that larval stages reflect the interrelationships of the groups; thus, because sea urchins and brittle stars have pluteus larvae, they form a natural group, and starfishes and sea cucumbers form another for the same reason. Some biochemical studies support this scheme. On the other hand, comparative anatomy and some paleontology studies suggest that brittle stars and starfishes may have originated from a crinoidlike ancestor and should be placed together, and their general star shape would support this. Modern sea cucumbers and sea urchins share a globoid body but little else; however, some fossil sea urchins with overlapping skeletal plates share several features with some sea cucumbers.

CLASSIFICATION

Distinguishing taxonomic features. The classification of the echinoderms underwent a great upheaval during the

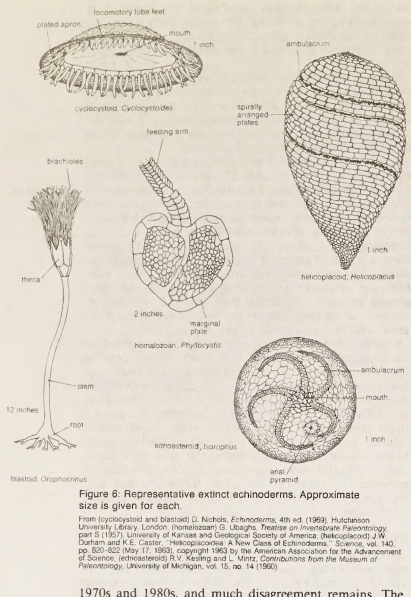


Figure 6: Representative extinct echinoderms. Approximate size is given for each.

From (cyclocystoid and blastoid) D. Nichols, *Echinoderms*, 4th ed. (1969), Hutchinson University Library, London. (homalozoan) G. Ubaghs, *Treatise on Invertebrate Paleontology*, part S (1967), University of Kansas and Geological Society of America. (heliocystoid) J.W. Durham and K.E. Caster, "Helocystocoidea: A New Class of Echinoderms," *Science*, vol. 140, pp. 820–822 (May 17, 1963), copyright 1963 by the American Association for the Advancement of Science. (edrioasteroid) R.V. Kesling and L. Mintz, *Contributions from the Museum of Paleontology*, University of Michigan, vol. 15, no. 14 (1960).

1970s and 1980s, and much disagreement remains. The five subphyla presented here are based upon combinations of characters: Homalozoa are asymmetrical; Blastozoa are stalked, with simple feeding apparatus; Crinozoa are stalked, with complex feeding apparatus; Asterozoa are star-shaped; Echinozoa are globoid to discoid. Below the subphylum level, the criteria for classification vary, but the skeleton is the most important; most groups can be characterized on the basis of skeletal characters alone.

Annotated classification. The echinoderms once were divided into two great groups, the Pelmatozoa and the Eleutherozoa. the names referring to living habits; pelmatozoans were attached to the seafloor for at least part of their life cycle while eleutherozoans were unattached animals capable of moving freely over the seafloor. It has been argued that such a separation is confusing, because each group contains a mixture of subgroups bearing no relationship to the evolutionary history of the phylum. The terms pelmatozoan and eleutherozoan are often used to describe the life habits of echinoderms. Some sea cucumbers, for example, have adopted a pelmatozoan habit, attaching themselves to rocks and feeding on plankton; others are eleutherozoan, moving about the seafloor while feeding, or even actively swimming.

The classification presented here is based upon current research by paleontologists and zoologists. Totally extinct classes, marked with a dagger (†), are known only as fossils.

PHYLUM ECHINODERMATA (echinoderms)

Marine invertebrates worldwide in distribution; skeleton composed of calcium carbonate in the form of calcite; most fossils and all living representatives with 5-part body symmetry (pentamerous); part of body cavity (coelom) comprises a water-vascular system. Cambrian to Recent. About 6,000 extant species, about 13,000 extinct species described.

†Subphylum Homalozoa (carpoids)

Middle Cambrian to Middle Devonian about 365,000,000–570,000,000 years ago; without 5-part symmetry; with fundamentally asymmetrical flattened body.

†Class Stylophora

Middle Cambrian to Upper Ordovician about 460,000,000–540,000,000 years ago; with unique single feeding arm sometimes interpreted as a stem.

†Class Homostelea

Middle Cambrian about 540,000,000 years ago; no feeding arm, but with stem of essentially 2 series of plates.

†Class Homoiostelea

Upper Cambrian to Lower Devonian about 400,000,000–510,000,000 years ago; with a feeding arm and a complex stem composed in part of more than 2 series of plates.

†Class Ctenocystoidea

Middle Cambrian about 540,000,000 years ago; no feeding arm and no stem, but with unique feeding apparatus consisting of a grill-like array of movable plates around mouth.

†Subphylum Blastozoa (blastozoans)

Cambrian to Permian about 280,000,000–540,000,000 years ago. Stalked echinoderms with soft parts enclosed in a globular theca (chamber) equipped with simple, erect food-gathering appendages (brachioles).

†Class Eocrinoidea

Lower Cambrian to Silurian about 430,000,000–570,000,000 years ago; body usually consisting of stem, theca, and feeding brachioles.

†Class Blastoidea

Silurian to Permian about 280,000,000–430,000,000 years ago; stem, theca with 18–21 plates arranged in 4 rings; numerous feeding brachioles; distinctive infoldings of theca (hydrospires) well developed.

†Class Paracrinoidea

Middle Ordovician about 460,000,000 years ago; with stem, theca, and arms with barblike structures (pinnules); plates of theca with pore system of unique type.

†Class Parablastoidea

Lower to Middle Ordovician about 460,000,000–500,000,000 years ago; resemble Blastoidea but differ in structure of ambulacra and in numbers of thecal plates.

†Class Rhombifera

Lower Ordovician to Upper Devonian about 350,000,000–500,000,000 years ago; theca globular; respiratory structures rhomboid sets of folds or canals.

†Class Diploporita

Lower Ordovician to Lower Devonian about 400,000,000–500,000,000 years ago; theca globular; respiratory structures pairs of pores.

Subphylum Crinozoa

Both fossil and living forms (Lower Ordovician about 500,000,000 years ago to Recent); with 5-part symmetry; soft parts enclosed in theca, which gives rise to 5 or more complex feeding arms.

Class Crinoidea (sea lilies and feather stars)

Lower Ordovician about 500,000,000 years ago to Recent; with, or secondarily without, a stem; theca reduced to small, cup-carrying hollow, usually branching, feeding arms with numerous small pinnules; includes fossil subclasses Camerata, Inadunata, and Flexibilia; living subclass Articulata, which includes stalked sea lilies and unstalked feather stars; about 700 living species.

Subphylum Asterozoa

Fossil and living forms (Lower Ordovician about 500,000,000 years ago to Recent); radially symmetrical with more or less star-shaped body resulting from growth of arms in 1 plane along 5 divergent axes; central mouth; 5 arms; dorsal tube feet and mouth.

Class Stelleroidea

Features as subphylum above.

†Class Somasteroidea

Lower Ordovician to Upper Devonian about 350,000,000 years ago. Superficially like Asteroidea, without a groove for tube feet.

Class Asteroidea (starfishes or sea stars)

Fossil and living forms (Middle Ordovician about 460,000,000 years ago to Recent); about 1,800 living species; arms broad, hollow; pinnate structure or arrangement of arms disrupted by dominant longitudinal growth axes; tube feet numerous, carried in grooves on the oral surface of the body; tube feet pointed or equipped with terminal suckers; respiration often by interradial gills on aboral surface of body; includes living orders Platyasterida, Paxillosida, Valvatida, Spinulosida, Forcipulatida, Notomyotida, and Brisingida.

Class Ophiuroidea (brittle stars or serpent stars)

Fossil and living forms (Ordovician about 460,000,000 years ago to Recent); disk sharply distinct from long, slender, solid arms; no furrow for tube feet; no suctorial tube feet; no anus; no pedicellariae; respiration by interradial gills on oral surface of body; includes living orders Oegophiurida, Phrynophiurida, and Ophiurida; about 2,000 living species.

Class Concentricycloidea (sea daisies)

Body flattened, disk-shaped, without obvious arms; water-vascular system with tube feet on oral surface of body; water-vascular canals form double ring; includes order Peripodida; 2 living species.

Subphylum Echinozoa

Fossil and living forms (Lower Cambrian about 570,000,000 years ago to Recent); radially symmetrical with fundamentally globoid body secondarily cylindrical or discoid; outspread arms or brachioles totally absent.

†*Class Cyclocystoidea*

Middle Ordovician to Middle Devonian about 375,000,000–460,000,000 years ago; small, disk-shaped; theca composed of numerous plates; ambulacral system with multiple branching.

†*Class Edrioasteroidea*

Lower Cambrian to Lower Carboniferous about 340,000,000–570,000,000 years ago; discoid to cylindrical; 5 well-developed straight or curved ambulacral food grooves radiate from a central mouth.

†*Class Edrioblastoidea*

Middle Ordovician about 375,000,000 years ago; stalked form with spheroidal theca; 5 well-developed food grooves.

†*Class Helicoplacoidea*

Lower Cambrian about 570,000,000 years ago; pear-shaped or spindle-shaped body with many plates arranged spirally.

†*Class Ophiocistioidea*

Lower Ordovician to Upper Silurian about 395,000,000–500,000,000 years ago; dome-shaped body partly or completely covered by well-developed test; 5 ambulacral tracts carry plated tube feet relatively enormous in size.

Class Holothuroidea (sea cucumbers)

Fossil and living forms (Ordovician about 460,000,000 years ago to Recent); cylindrical body, elongated orally–aborally, with mouth at or near one end, anus at or near the other; mouth surrounded by conspicuous ring of feeding tentacles; no spines or pedicellariae; single interradial gonad; skeleton usually reduced to form microscopic spicules; includes living orders Dendrochirotida, Dactylochirotida, Aspidochirotida, Elasipodida, Molpadiida, and Apodida; 1,100 living species.

Class Echinoidea (sea urchins, heart urchins, sand dollars)

Fossil and living forms (Ordovician 460,000,000 years ago to Recent); globular, discoid, or oval in shape, with complete skeleton (test) of interlocking plates bearing movable spines and pedicellariae; mouth directed downward; anus present; 5 or fewer interradial gonads. Includes subclass Perischoechinoidea with living order Cidaroida, and subclass Euechinoidea with living superorders Diadematacea and Echinacea (comprising the "regular" echinoids), and Gnathostomata and Atelostomata (comprising the "irregular" echinoids); 900 living species.

Critical appraisal. No classification satisfies everyone, and this is especially true for the echinoderms. Some modern scientists argue that fossils contribute little to our understanding of the interrelationships of living groups because fossil forms are different from recent forms and because many of the forms that link the groups in a classification scheme are missing. They believe instead that the higher classification of the echinoderms should be based upon a study of the embryology and anatomy of living groups and that the fossil groups should be inserted wherever they best seem to fit. In other classification schemes, scientists regard the fossils as a logical starting point—the "roots of the tree." A classification proposed in the early 1960s based upon growth patterns enjoyed wide acceptance until further research showed numerous flaws in the overall scheme. The current classification is decidedly a "hybrid," incorporating data from several fields of biologic research.

It has been strongly argued that some members of the subphylum Homalozoa are true chordates rather than echinoderms. This theory has not received wide acceptance. If it eventually proves to be correct, a drastic reevaluation of the Echinodermata would be required. The phylum would then share with chordates a latticelike calcite skeleton and a water-vascular system.

New discoveries and new theories are continually reshaping the classification. The subphylum Blastozoa, proposed in the early 1970s, has gained wide acceptance. The extinct classes Helicoplacoidea and Ctenocystoidea were suggested in the 1960s, and their discovery caused a reassessment of the classification of the phylum. Extinct classes Lepidocystoidea and Camptostromatoidea have been eliminated and their members distributed among other echinoderm groups. The extant class Concentricycloidea was described in 1986 and is the first new class of living echinoderms to be named since 1821. Some argue that the concentricycloids are extreme forms of starfish that properly belong in the class Asteroidea. Less systematic importance is attached to the characters that are regarded as of class rank.

BIBLIOGRAPHY. RICHARD A. BOOLOOTIAN (ed.), *Physiology of Echinodermata* (1966), a comprehensive survey of biology and physiology; AILSA M. CLARK, *Starfishes and Related Echinoderms*, 3rd ed. (1977), an introductory work, with emphasis on living forms, classification, and biology; LIBBIE HENRIETTA HYMAN, *The Invertebrates*, vol. 4, *Echinodermata. The Coelomate Bilateria* (1955), a classic survey of anatomy and biology; MICHEL JANGOUX and JOHN M. LAWRENCE (eds.), *Echinoderm Nutrition* (1982), a thorough survey of feeding biology, and *Echinoderm Studies* (1983), a collection of review articles on all aspects of echinoderm biology; JOHN M. LAWRENCE, *A Functional Biology of Echinoderms* (1987), a discussion of echinoderm food acquisition, reproduction, and other aspects of their lives; RAYMOND C. MOORE and CURT TEICHERT (eds.), *Treatise on Invertebrate Paleontology*, pt. S, T, and U, "Echinodermata" (1966–78), a detailed treatment; DAVID NICHOLS, *Echinoderms*, 4th ed. (1969), a general work, including a treatment of fossil forms; and ANDREW SMITH, *Echinoid Palaeobiology* (1984), a study of anatomy and paleontology of sea urchins. VICKI PEARSE *et al.*, *Living Invertebrates* (1987), includes a well-illustrated discussion of the living echinoderms. For an exposition at a less advanced level, see RALPH BUCHSBAUM *et al.*, *Animals Without Backbones*, 3rd ed. (1987).

(D.L.P./J.E.M.)

Eclipse, Occultation, and Transit

Eclipse, occultation, and transit are related astronomical events in which three celestial bodies are aligned; the body passing between the other two obscures some or all of the light of one of these as seen from the other. The difference in terminology is related to the relative sizes of the obscuring and obscured bodies.

The Sun is eclipsed when the Moon comes between it and the Earth; the Moon is eclipsed when it moves into the shadow of the Earth cast by the Sun. Eclipses of natural or artificial satellites of a planet occur as the satellites move into the planet's shadow. The two component stars of an eclipsing binary star move around each other in such a way that their orbital plane passes through or very near the Earth, and each star periodically eclipses the other as seen from the Earth.

When the apparent size of the eclipsed body is much smaller than that of the eclipsing body, the phenomenon is known as an occultation. Examples are the disappearance of a star, nebula, or planet behind the Moon, or the vanishing of a natural satellite or space probe behind some body of the solar system.

A transit occurs when, as viewed from the Earth, a relatively small body passes across the disk of a larger body, usually the Sun or a planet, eclipsing only a very small area: Mercury and Venus periodically transit the Sun, and a satellite may transit its planet.

For coverage of related topics in the *Macropædia* and *Micropædia,* see the *Propædia,* sections 132 and 133, and the *Index.*

This article is divided into the following sections:

Phenomena observed during eclipses 866
  Lunar eclipse phenomena
  Solar eclipse phenomena
The geometry of eclipses, occultations, and transits 867
  Eclipses of the Sun
  Eclipses of the Moon
  Eclipses, occultations, and transits of satellites
The frequency of solar and lunar eclipses 868
  Cycles of eclipses
  Prediction and calculation of solar and lunar eclipses
Eclipse research activities 870
  Solar research
  Lunar research
Transits of Mercury and Venus 870
Occultations by the Sun and Moon 871
Eclipsing binary stars 872
Eclipses in history 872
  Literary and historical references
  Uses of eclipses for chronological purposes
  Uses of eclipses for astronomical purposes
Bibliography 877

PHENOMENA OBSERVED DURING ECLIPSES

Lunar eclipse phenomena. The Moon may, when full, enter the shadow of the Earth. The motion of the Moon around the Earth is from west to east relative to the Sun, so that, for an observer facing south, the shadowing of the Moon begins at its left edge (if the Moon were north of the observer, as, for example, in parts of the Southern Hemisphere, the opposite would be true). If the eclipse is a total one and circumstances are favourable, the Moon will pass through the umbra, or darkest part of the shadow, in about two hours (see Figure 1). During this time, the Moon is usually not quite dark. A part of the sunlight, especially the redder light, penetrates the Earth's atmosphere, is refracted into the shadow cone, and reaches the Moon. Meteorologic conditions on Earth strongly affect the amount and colour of light that can penetrate the atmosphere. Generally, the totally eclipsed Moon is clearly visible and has a reddish-brown, coppery colour, but the brightness varies strongly from one eclipse to another.

Total lunar eclipse

After the Moon leaves the umbra, it must still pass through the penumbra of the shadow. When the border between umbra and penumbra is visible on the Moon, the border is seen to be part of a circle, the projection of the circumference of the Earth. This is a direct proof of the spherical shape of the Earth. Because of the Earth's atmosphere, the edge of the umbra is rather diffuse, and the times of contact between the Moon and the umbra cannot be observed accurately.

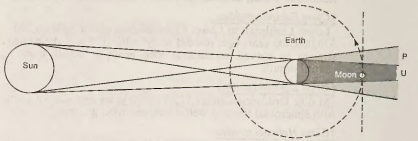


Figure 1: *Eclipse of the Moon.*
The Moon revolving in its orbit around the Earth passes through the shadow of the Earth. U (umbra) is the total shadow, P (penumbra) the partial shadow.

Cooling of the lunar surface during eclipses

During the eclipse, the surface of the Moon cools at a rate dependent on the constitution of the lunar soil, which is not everywhere the same. Many spots on the Moon sometimes remain brighter than their surroundings during totality—particularly in their output of infrared radiation—possibly because their heat conductivity is less, but the cause is not fully understood.

An eclipse of the Moon can be seen under similar conditions at all places on the Earth where the Moon is above the horizon.

Solar eclipse phenomena. Totality at any particular solar eclipse can only be seen from a relatively narrow belt on Earth. The various phases observable at a total solar eclipse are illustrated in Figure 2A. "First contact" designates the moment when the disk of the Moon, invisible against the bright sky background, just touches the disk of the Sun. The partial phase of the eclipse then begins, as a small indentation in the western rim of the Sun becomes noticeable. The dark disk of the Moon now gradually moves across the Sun's disk, and the bright area of the Sun is reduced to a crescent. The sunlight, shining through gaps in foliage and other small openings, is then seen to form little crescents of light that are images of the light source, the Sun. Toward the beginning of totality, the direct light from the Sun diminishes very quickly and the colour changes. The sky becomes dark, but, along

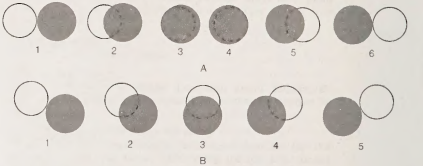


Figure 2: *Successive phases of a solar eclipse.*
The dark disk of the Moon gradually moves across the disk of the Sun from west (right) to east (left). (A) Total eclipse: (1) first contact; (2) partial phase; (3) second contact, beginning of totality; (4) third contact, end of totality; (5) partial phase; (6) fourth contact. (B) Partial eclipse: (1) first contact; (3) maximum phase; (5) last contact.

the horizon, the Earth's atmosphere still appears bright because the umbra of the Moon's shadow on the Earth extends over a rather narrow region. The scattered light coming in from a distance beyond this region produces weird effects. Men, birds, and other animals react with fear; birds may go to roost as they do at sunset.

Baily's beads

As the tiny, narrow crescent of sunlight disappears, little bright specks remain where depressions in the Moon's edge, the limb, are last to obscure the Sun's limb. These specks are known as Baily's beads, after the 18th-century English astronomer Francis Baily, who first drew attention to them. The beads vanish at the moment of second contact, when totality sets in. This is the climax of the eclipse. The reddish prominences and chromosphere of the Sun, around the Moon's limb, can now be seen. The brighter planets and stars become visible in the sky. The white corona extends out from the Sun to a distance greater than the Sun's diameter, at which point it fades completely. The temperature in the path of totality falls by some degrees. The light of totality is much brighter than that of the Full Moon but is quite different (see below).

The moment of third contact approaches, at which time many of the phenomena of second contact appear again in reverse order. Suddenly the first Baily's bead appears, now on the other side of the Moon. More beads of light follow, the Sun's crescent grows again, the corona disappears, daylight brightens, and the stars and planets fade from view. The thin crescent of the Sun gradually widens, and about one and a quarter hours later the eclipse ends with fourth contact, when the last encroachment made by the Moon on the Sun's rim disappears.

During the partial phase, both before and after totality, it is absolutely essential to protect the eyes against injury by the intense brilliance of the Sun. It should never be viewed directly except through strong filters, a dark smoked glass, or a heavily fogged photographic plate or film.

Shadow bands

When totality is imminent and only a small crescent of the Sun remains, the so-called shadow bands can often be seen on plain light-coloured surfaces, such as open floors and walls. These are striations of light and shade, moving and undulating, several centimetres (or inches) wide. Their velocity and direction depend on air currents at various heights, as they are caused by refraction of sunlight by small inhomogeneities in the Earth's atmosphere. A similar phenomenon is the projection of water waves on the bottom of a sunlit swimming pool or bath.

THE GEOMETRY OF ECLIPSES, OCCULTATIONS, AND TRANSITS

Eclipses of the Sun. An eclipse of the Sun takes place when the Moon comes between the Earth and the Sun so that the Moon's shadow sweeps over the face of the Earth (see Figure 3). This shadow consists of two parts; the umbra, or total shadow, a cone into which no direct sunlight penetrates; and the penumbra, or half shadow, which is reached by light from only a part of the Sun's disk.

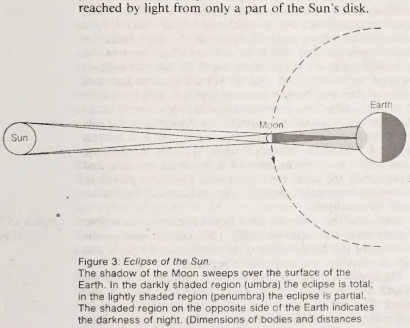


Figure 3: *Eclipse of the Sun.*
The shadow of the Moon sweeps over the surface of the Earth. In the darkly shaded region (umbra) the eclipse is total; in the lightly shaded region (penumbra) the eclipse is partial. The shaded region on the opposite side of the Earth indicates the darkness of night. (Dimensions of bodies and distances are not to scale.)

Total solar eclipse

To an observer within the umbra, the Sun's disk appears completely covered by the disk of the Moon; such an eclipse is called total. To an observer within the penumbra, the Moon's disk appears projected against the Sun's disk so as to overlap it partly; the eclipse is then called partial for that observer. The umbra cone is narrow at the distance of the Earth, and a total eclipse is observable only within the narrow strip of land or sea over which the umbra passes. A partial eclipse may be seen from places within the large area covered by the penumbra. Sometimes the Earth intercepts the penumbra of the Moon but is missed by its umbra; only a partial eclipse of the Sun is then observed anywhere on the Earth.

By a remarkable coincidence, the sizes and distances of the Sun and Moon are such that they appear as very nearly the same angular size (about 0.5°) at the Earth, but their apparent sizes depend on their distances from the Earth. The Earth revolves around the Sun in an elliptical orbit, so that the distance of the Sun changes slightly during a year, with a correspondingly small change in the apparent size, the angular diameter, of the solar disk. In a similar way, the apparent size of the Moon's disk changes somewhat during the month because the Moon's orbit is also elliptical. When the Sun is nearest to the Earth and the Moon is at its greatest distance, the apparent disk of the Moon is smaller than that of the Sun. If an eclipse of the Sun occurs at this time, the Moon's disk passing over the Sun's disk cannot cover it completely but will leave the rim of the Sun visible all around it. Such an eclipse is said to be annular. Total eclipses and annular eclipses are called central.

In a partial eclipse (Figure 2B), the centre of the Moon's disk does not pass across the centre of the Sun. After the first contact, the visible crescent of the Sun decreases in width until the centres of the two disks reach their closest approach. This is the moment of maximum phase, and the extent is measured by the ratio between the smallest width of the crescent and the diameter of the Sun. After maximum phase, the crescent of the Sun widens again until the Moon passes out of the Sun's disk at the last contact.

Eclipses of the Moon. When the Moon moves through the shadow of the Earth (see Figure 1) it loses its bright direct illumination by the Sun, although its disk still remains faintly visible. As the shadow of the Earth is directed away from the Sun, a lunar eclipse can occur only at the time of Full Moon—that is, when the Moon is on the side of the Earth opposite to that of the Sun. A lunar eclipse appears much the same at all points of the Earth from which it can be seen. When the Moon enters the penumbra, a penumbral eclipse occurs. The dimming of the Moon's illumination by the penumbra is so slight as to be scarcely noticeable, and penumbral eclipses are rarely watched. After a part of the Moon's surface is in the umbra and thus darkened, the Moon is said to be in partial eclipse. After about an hour, when the whole disk of the Moon is within the umbra, the eclipse becomes total. If the Moon's path leads through the centre of the umbra, the total eclipse can be expected to last about an hour and three-quarters.

Penumbral eclipse of the Moon

Eclipses, occultations, and transits of satellites. These phenomena are conveniently illustrated by the four largest satellites of Jupiter, whose eclipses provide a frequently occurring and fascinating spectacle to the telescopic observer. The three innermost moons (Io, Europa, and Ganymede) disappear into the shadow of Jupiter at each revolution, though the fourth (Callisto) is not eclipsed every time. Because of the sizable dimensions of these bodies, some minutes elapse between first contact with the shadow and totality. The orbits of these satellites lie nearly in the same plane as Jupiter's orbit around the Sun, and, at practically every revolution of each satellite, the following four eclipse phenomena take place: (1) eclipse of the satellite when it passes through Jupiter's shadow; (2) occultation of the satellite when it disappears behind the planet, as seen from the Earth; (3) transit of the satellite across the disk of Jupiter; and (4) transit of the shadow of the satellite across the planet's disk.

Figure 4 illustrates these phenomena; it shows Jupiter

and the orbit of one of its satellites, the direction of the sunlight illuminating the system, and the direction toward the Earth, from where the observation is made. When the satellite arrives at the point S_1 of its orbit, it enters Jupiter's shadow (eclipse) and vanishes. At S_2 it comes out of the shadow, but, to the terrestrial observer, it is now hidden behind the planet (occultation) until at S_3 it reappears at the limb. When the satellite reaches the position S_4, its shadow falls on Jupiter, causing a small dark spot on its surface. Seen from the Earth, the satellite is to the left of Jupiter approaching Jupiter's limb, at the time that its shadow spot passes across the planet's disk (transit of shadow). At S_5 the satellite starts to pass in front of the planet (transit of satellite), following its shadow spot. Both Jupiter and the satellite must have their illuminated sides facing the Earth. They differ little in total surface brightness; near the limb the satellite is somewhat brighter than the planet's surface on which it appears projected, but near the middle of the disk it is hardly distinguishable. At S_6 the shadow leaves the planet, and at S_7 the satellite emerges at the limb.

Proof of finite speed of light

Historically, the eclipses of Jupiter's satellites are important, for they provided one of the earliest proofs of the finite speed of light. It is possible to calculate with considerable precision the times of disappearance and reappearance of a satellite undergoing eclipse. The Danish astronomer Ole Rømer in 1675 noticed discrepancies between the observed and calculated times of such eclipses, which he correctly explained as being due to the difference in the travel time of light when the Earth is nearest to Jupiter or farther away from it.

A related phenomenon is the occultation of a space probe by a planet. During the beginning and the end of such an occultation, signals sent out by the spacecraft and received on Earth have penetrated the planet's atmosphere and can yield information about atmospheric density and composition.

On March 10, 1977, the planet Uranus passed between the Earth and a bright star. The event was observed by several teams of astronomers, who hoped to derive an accurate estimate of the diameter of the planet from their data. To their surprise, however, the light from the star was briefly obscured several times before and after the disk of Uranus occulted it. It was concluded that Uranus has a system of rings somewhat like those of Saturn.

THE FREQUENCY OF SOLAR AND LUNAR ECLIPSES

A solar eclipse, especially a total one, can be seen from only a limited part of the Earth, while the eclipsed Moon can be seen at the time of the eclipse wherever the Moon is above the horizon.

In most calendar years there are two lunar eclipses; in some years one or three or none occur. Solar eclipses occur two to five times a year, five being exceptional; there were five in 1935 and will be again in 2206. The average number of total solar eclipses in a century is 66 for the Earth as a whole.

Numbers of solar eclipses predicted to take place during the 20th to the 25th century are:

1901–2000:	228	eclipses,	of which	145	are	central
2001–2100:	224	"	"	144	"	"
2101–2200:	235	"	"	151	"	"
2201–2300:	248	"	"	156	"	"
2301–2400:	248	"	"	160	"	"
2401–2500:	237	"	"	153	"	"

Any point on Earth may, on the average, experience no more than one total solar eclipse in three to four centuries. The situation is quite different for lunar eclipses. An observer remaining at the same place (and granted cloudless skies) can see 19 or 20 lunar eclipses in 18 years: three or four total eclipses and six or seven partial eclipses may be visible from beginning to end, and five total eclipses and four or five partial eclipses at least partially visible.

Duration of eclipses

All these numbers can be worked out from the geometry of the eclipses. A total lunar eclipse can last as long as an hour and three-quarters, but for a solar total eclipse maximum duration of totality is only $7^1/_2$ minutes. This difference results from the fact that the Moon is much smaller in cross section than the extension of the Earth's shadow but can be only a little greater in apparent size than the Sun.

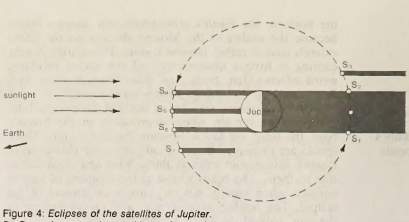


Figure 4: *Eclipses of the satellites of Jupiter.* S_1–S_7 mark successive positions of one of these satellites as it revolves in its orbit around Jupiter.

Cycles of eclipses. The eclipses of the Sun and Moon occur at New Moon and Full Moon, respectively, so that one basic time period involved in the occurrence of eclipses is the synodic month; *i.e.*, the time of one revolution of the Moon around the Earth with respect to the Sun.

Eclipses and the synodic month

A solar eclipse does not occur at every New Moon, because the Moon's orbit plane is inclined to the ecliptic, the plane of the orbit of the Earth around the Sun. The angle between the planes is about 5°; thus the Moon can pass well above or below the Sun. The line of intersection of the planes is called the line of the nodes, being the two points where the Moon's orbit intersects the ecliptic plane. The ascending node is the point where the Moon crosses the ecliptic from south to north, and the descending node that where it goes from north to south. The nodes move along the orbit from west to east, going completely around the ecliptic in about 19 years. The Moon's revolution from one node to the same node again (called the draconic month, 27.212220 days) takes somewhat less time than a revolution from Full Moon to Full Moon (the synodic month, 29.530589 days). For an eclipse to occur, the Moon has to be near one of the nodes of its orbit. The draconic month, therefore, is the other basic period of eclipses.

The saros period

Resonance between these two periods produces what is called the saros period, after which time Moon and Sun return very nearly to the same relative positions. The saros was known to the ancient Babylonians (see below *Use of eclipses for astronomical purposes: In ancient astronomy*). It comprises 223 lunations (the time period from New Moon to New Moon)—that is, 6,585.321124 days, or 241.9986 draconic months. This is nearly a whole number, so that the Full Moon is in almost the same position (*e.g.*, very near a node) at the beginning and end of a saros, a lapse of time equalling 18 years and $11^1/_3$ days or $10^1/_3$ days if five leap years fall within the period. Thus, there is usually a close resemblance between any eclipse and the one taking place 18 years and 11 days earlier or later. Since the date differs by only about 11 days in the calendar year, the latitudes on Earth of the two eclipses will be about the same, and so will the relative apparent sizes of Sun and Moon. The saros period also comprises 238.992 anomalistic months, again nearly a whole number. In one anomalistic month, the Moon describes its orbit from perigee to perigee, in which point it is nearest to the Earth. So the Moon's distance from the Earth is the same after a whole number of anomalistic months and very nearly the same after one saros. The saros period is, therefore, extremely useful for the prediction of both solar and lunar eclipses.

Because of the extra one-third day in the saros, the eclipse recurs each time approximately 120° farther west on the surface of the Earth. After three saroses, or 54 years and about a month, the longitude is repeated.

There is a regular shift on Earth to the north or to the south of successive eclipse tracks from one saros to the next. The eclipses occurring when the Moon is near its ascending node shift to the south, those happening when it is near its descending node shift to the north. A saros

series of eclipses begins its life at one pole of the Earth and ends it at the other. Every saros series lasts about 1,300 years and comprises 73 eclipses. About 42 of these series overlap at any time.

The inex period

Two consecutive saros series are separated by the inex period, 29 years minus 20 days—that is, 358 synodic months—after which time the New Moon has come from one node to the opposite node. The lifetime of an inex group is about 23,000 years, 70 groups coexisting, each comprising 780 eclipses. All other cycles in eclipses are combinations of the saros and the inex.

Prediction and calculation of solar and lunar eclipses. The problem may be divided into two parts. The first is to find out when an eclipse will occur, the other to determine when and where it will be visible.

For this purpose it is convenient first to consider the Earth as fixed and to suppose the observer looking out from its centre. To this observer, O in Figure 5, the Sun and Moon appear projected on the celestial sphere. While this sphere appears to him to rotate daily, as measured by the positions of the stars, around the line PP′ (the Earth's axis of rotation), the Sun's disk, S, appears to travel slowly along the great circle EE′ (the ecliptic), making a complete revolution in one year. At the same time the Moon's disk, M, revolves along the circle LL′ once during a lunar month. The angular diameters of the two disks S and M are each about 0.5° but vary slightly.

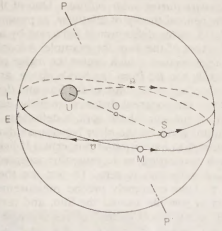


Figure 5: Apparent motions of the Sun and the Moon on the celestial sphere (see text).

Every month the Moon's disk revolving along LL′ will overtake the more slowly moving Sun once, at the moment of New Moon. Usually the Moon's disk will pass above or below the Sun's disk. Overlapping of the two results in an eclipse of the Sun, which can happen only when the New Moon occurs at a moment when the Sun is near the points ☊ or ☋; these signs denote the ascending and descending nodes of the Moon's orbit.

The crosscut of the umbra, the shadow cone of the Earth, at the distance of the Moon (as shown in Figure 5), may be projected like a disk U onto the celestial sphere. It subtends an angle of about 1.4°; its centre will always be opposite to the Sun's disk and travel along EE′. A lunar eclipse occurs whenever the Moon's disk overlaps the shadow disk; this happens only when the shadow disk is near one of the nodes or the Sun is near the opposite node. The Sun's passage through the lunar nodes is thus the critical time for both solar and lunar eclipses. The Moon's orbit plane, represented by the circle LL′, is not fixed, and its nodes move slowly along the ecliptic in the direction indicated by the arrow, making a complete revolution in about 19 years. The interval between two successive passages of the Sun through one of the nodes is termed an "eclipse year," and, since the Moon's node moves so as to meet the advancing Sun, this interval is about 18.6 days less than a tropical (or ordinary) year.

The eclipse year

In Figure 6 the region of the ascending node as seen from the centre of the sphere is much enlarged. Here the node is kept fixed and the apparent motions of the Sun and the Moon are shown relative to the node. To the imaginary observer at the centre of the Earth, the Sun's disk will travel along the circle EE′, the Moon's disk along

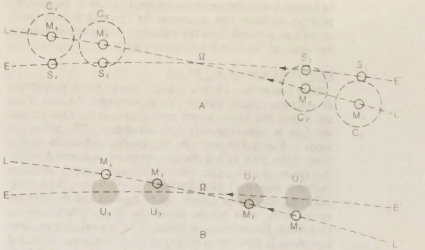


Figure 6: *Ascending node of the Moon's orbit as seen from the centre of the sphere.*
Conditions necessary for (A) a solar eclipse and (B) a lunar eclipse (see text).

LL′. The Sun is so distant compared with the size of the Earth that, from all places on the Earth's surface, the Sun is seen nearly in the same position as from the centre. But the Moon is relatively near and its projected position on the celestial sphere is different for various observing stations on the Earth; it may be displaced as much as 1° from the position in which it is seen from the centre of the Earth. If the radius of the Moon's disk is enlarged by 1°, a circle, C, is obtained that encloses all possible positions of the Moon's disk seen from anywhere on the Earth. Conversely, if any circle of the Moon's size is drawn inside this "Moon circle," C, there is a place on the Earth from which the Moon is seen in that position.

Accordingly, an eclipse of the Sun occurs somewhere on Earth whenever the Moon overtakes the Sun in such a position that the Moon circle, C, passes over the Sun's disk; when the latter is entirely covered by the Moon circle, the eclipse will be central (*i.e.*, total or annular). From Figure 6A, it is evident that a solar eclipse will take place if a New Moon occurs while the Sun moves from S_1 to S_4. This period is the eclipse season; it starts 19 days before the Sun passes a node and ends 19 days thereafter. Since there is a New Moon every month, at least one solar eclipse, and occasionally two, occurs during every eclipse season—of which there are two in each calendar year. A fifth solar eclipse during a calendar year is possible because part of a third eclipse season may occur at the beginning of January or at the end of December.

The eclipse season

Figure 6B illustrates the condition necessary for a lunar eclipse. If a Full Moon occurs within 13 days of a node passage of the Sun (when the shadow disk, U, passes the ascending node), the Moon will be eclipsed. Most eclipse seasons, but not all, will thus also contain a lunar eclipse. When three eclipse seasons fall in a calendar year, there may be three lunar eclipses in that year. Eclipses of the Sun are evidently more frequent than those of the Moon. Solar eclipses, however, can only be seen from a very limited region of the Earth, whereas lunar eclipses are visible from an entire hemisphere.

During a solar eclipse the shadow cones (umbra and penumbra) of the Moon sweep across the face of the Earth (Figure 3), while at the same time the Earth is rotating on its axis. Within the narrow area covered by the umbra, the eclipse is total. Within the wider surrounding region covered by the penumbra, the eclipse is partial.

The astronomical ephemerides, or tables, published for each year provide maps tracing the paths of the more important eclipses in considerable detail, as well as data for accurate calculation of the times of contact at any given observing station. Calculations are made some years ahead in Ephemeris Time (ET), which is defined by the orbital motion of the Earth and the other planets. At the time of the eclipse, the correction is made to Universal Time (UT), which is defined by the rotation of the Earth and is not rigorously uniform.

Accurate short-range predictions

It is possible with the aid of modern tables to accurately predict solar eclipses several years ahead. For predictions

of longer range, the main uncertainty is that of the Moon's motion. Eclipses can of course be "predicted backward" as well as forward, and the calculation of ancient eclipses has been of value in historical research.

ECLIPSE RESEARCH ACTIVITIES

Solar research. During a solar eclipse, the Moon serves as a screen outside the Earth's atmosphere. As seen from the Earth, the Moon's dark projection on the Sun crosses the Sun at approximately 350 kilometres (220 miles) per second. For a few seconds, the bright, ordinarily visible disk of the Sun, called the photosphere, is eclipsed, while the chromosphere, the lower solar atmosphere, remains visible at the Sun's edge. During this brief time, light emitted from the chromosphere can be studied and its decrease in intensity with height above the photosphere can be measured. Discrimination between layers within the chromosphere is possible to some extent. An instrument called the coronagraph has been developed to obscure the brilliant photosphere artificially at times outside of eclipses, but the study of the thin layers of the solar chromosphere and corona is still best done during an eclipse.

The first photograph of a solar eclipse was taken in 1851, and the first of scientific importance by the British scientist and inventor Warren de la Rue and the Italian astronomer Angelo Secchi in 1860.

Spectroscopic observations. The ordinary spectrum of the Sun contains a brilliant multicoloured background—the continuum. This radiation emanates primarily from the lower layers of the solar atmosphere. Cooler atoms higher up, however, selectively absorb the radiations, so that the solar spectrum consists of the bright rainbow background with many narrow gaps—dark lines—where the light has been absorbed.

Spectroscopy was first applied to the eclipsed Sun in 1868, when the path of totality passed over India and Malaya and the spectrum of bright lines originating in the solar prominences was observed. The British astronomer Joseph Norman Lockyer thought that one of these spectral lines was emitted by an unknown chemical element, which he called helium (from Greek *hēlios,* "sun"); helium was not identified on Earth until 1895.

At the moment of totality, when the Moon obliterates the last trace of the bright photosphere, with its dark-line spectrum, the light from the upper tenuous layers of the solar atmosphere (which is usually lost in the much brighter photospheric light) flashes into view with the characteristic bright-line spectrum of a luminous gas. This spectrum, first observed in 1870, disappears within four to five seconds, as the Moon moves across the disk of the Sun. Because of its evanescent character, it is called the flash spectrum. A second flash occurs at the end of totality. Analysis of flash spectra has led to some surprising results. The spectrum matches the dark-line spectrum only roughly. The lines of the neutral metals are of comparable strength in the two spectra, but those of the ionized metals (*i.e.,* those made up of atoms that have lost one or more electrons) are markedly enhanced in the flash spectrum. The difference is attributed, in part, to lower pressures in the upper layers of the solar atmosphere, but high temperature appears to contribute to the increased excitation. This condition is especially true for the flash lines of ionized helium, which do not appear at all in the ordinary dark-line spectrum. Such flash lines require excitation temperatures of at least 25,000 K (about 45,000° F) for their production, whereas the temperature required to produce the observed quality and quantity of bright emission from the surface is only 6,000 K.

Flash-spectrum analysis

Other observations. Extending upward from the chromosphere, and closely related to it, are the so-called prominences, one of the striking features of a total eclipse, which project outward into space. They appear as rose-coloured patches of flame, projecting well beyond the limb of the Moon, and consist of long interlacing filaments of incandescent gas.

The corona of the Sun can be seen during totality. One of the most beautiful of natural phenomena, the solar corona shines like finely etched white frost against the deep blue of the eclipse-darkened sky. The form of the corona presented at different eclipses is almost infinitely variable. On occasion, usually when sunspots are near a minimum, long streamers extend four or five solar diameters away from the Sun. At other times, especially close to sunspot maximum, the corona is more nearly circular but with jagged, petallike extensions. The corona is faint, about 500,000 times less brilliant than the Sun itself. Consequently, the sky glare surrounding the Sun ordinarily hides the coronal details, and, before the development of the coronagraph, it was believed to be impossible to record the corona except during an eclipse.

Solar corona

Also interesting are observations of the radio emission of the Sun during a solar eclipse. When the Moon crosses different parts of the Sun, inhomogeneities in the generation of solar radio waves can be localized.

For geodetic purposes (measuring the Earth), the exact timing of the moments of the contacts of Moon and Sun are important, because these depend on the site of observation and so on the shape of the Earth.

The Sun's ultraviolet and X radiation help create and influence the ionosphere, a region in the Earth's upper atmosphere where many atoms are ionized. The ionosphere changes during an eclipse as the ionizing radiation is cut off. So that observations of the ionosphere made during an eclipse can be interpreted correctly, it is necessary to predict the times and phases of the eclipse as it reaches the high layers of the atmosphere, where they may be quite different from times and phases on the surface.

Tests of relativity during solar eclipse. One of the predictions of the general theory of relativity, as presented by Einstein in 1915, is the deflection of light rays by a gravitational field—that of the Sun, for example. According to this theory the deflection, which causes the image of a star to appear slightly too far from the Sun's image, amounts to 1.75 seconds of arc at the limb of the Sun and decreases in proportion to the apparent distance from the centre of the solar disk of the star whose light is deflected. This is twice the amount given by the older Newtonian dynamics if light is assumed to have inertial properties. If light does not have such properties, as is generally accepted now, the Newtonian deflection is zero. To test the theory, it is necessary to have extremely precise measurements of as many stars as possible around the Sun, and preferably close to it. The brightness of the corona prevents observation of stars closer to the Sun than one solar diameter. During totality the sky around the Sun is photographed with long-focus cameras to give the largest possible scale. The equipment is left in place, and about half a year later the same region of the sky with the same stars can be photographed again, this time during the night and without the Sun's disturbing gravity field. Comparison of the two sets of photographs shows the amount of gravitational displacement of starlight by the Sun and serves as a test of theory. Results of a number of eclipse observations made since 1918 have verified Einstein's prediction, though the deduced values are somewhat uncertain as a consequence of the small displacement.

Verification of Einstein's prediction

Lunar research. Lunar eclipses can yield information about the cooling of the Moon's soil when the Sun's radiation is suddenly removed and, therefore, about the soil's conductivity of heat and its structure. Infrared and radio-wavelength radiations from the Moon decline in intensity more slowly than does visible light emission during an eclipse because they are emitted from below the surface, and measurements indicate how far the different kinds of radiation penetrate into the lunar soil. Infrared observations show that at many "bright spots" the soil retains its heat much longer than in surrounding areas. Because of the absence of a lunar atmosphere, the solid surface is exposed to the full intensity of ultraviolet and particulate radiation from the Sun, which may give rise to fluorescence in some rock materials. Observations during lunar eclipses have given positive results for this phenomenon, with the appearance of abnormal bright regions in the obscured parts of the Moon.

TRANSITS OF MERCURY AND VENUS

At the time of inferior conjunction (*i.e.,* when moving between the Earth and the Sun) Mercury, seen from the

Earth, usually passes north or south of the Sun because of the inclination of its orbit. But if the conjunction occurs when Mercury is near one of the nodes of its orbit, the planet crosses the disk of the Sun as a small black circular spot, visible only with a telescope. Since the Earth passes Mercury's nodes on May 7 and November 9, transits can occur only near those dates. The shortest interval between two successive transits is seven years.

Transits of Venus can be seen without a telescope if the eyes are properly protected. When the transit is central, it takes about eight hours. The phenomenon is rare and can happen only within a day or two of the dates when the Earth passes the nodes of Venus' orbit—that is, on June 7 and December 8. The transits occur in pairs, with an interval of eight years between members of a pair; between the pairs, more than 100 years elapse.

Transits of Venus are helpful in finding the parallax and from it the distance of the Sun, as first pointed out by the British astronomer Edmond Halley in 1679. Parallax is the apparent difference in direction of an object when observed from different positions. The transits of June 1761 and 1769 and those of December in 1874 and 1882 were, thus, extensively observed. The next pair of transits of Venus are expected on June 8, 2004, and June 6, 2012.

One kind of observation for determining parallax consists in fixing the times of the contacts of the disks of the planet and the Sun from different points on the Earth. The observers at past transits became aware of a few remarkable phenomena. When Venus was partially overlapping the disk of the Sun, the part of the limb of the planet that extended beyond the Sun was seen to be surrounded with a radiant aureole, which observers of the transit in 1761 ascribed to the presence of an atmosphere on Venus. A second phenomenon was seen just after second and before third contact (see Figure 1), when Venus just touched the Sun's limb on the inside; this consisted in the development of a little dark connection—the so-called black drop—between Venus and the limb. Because of the black drop, the times of contact could not be sharply defined. Presumed causes of the black drop are diffraction, atmospheric agitation, and instrumental factors.

The black drop

The amount of sunlight intercepted during a transit depends on the diameter of the planet, and measuring this amount of sunlight may be one of the most accurate ways of determining the planet's diameter.

OCCULTATIONS BY THE SUN AND MOON

The occultation of the Crab Nebula by the solar corona, the extensive outer atmosphere of the Sun, takes place each year in June. The Crab Nebula is a radio source (Taurus A), and, when the occultation occurs, its pattern of radio emission is observed to broaden significantly. The broadening is attributed to scattering of radio-frequency radiation by the density irregularities in the extended corona. The nearest approach of Sun and Crab Nebula is at about five solar radii; the scattering is brought into evidence up to 60 solar radii.

Lunar occultation of stellar bodies

The Moon sometimes occults a planet (see Figure 7) but very often occults a star. As a consequence of the inclination of the Moon's orbit with respect to the ecliptic (the Sun's apparent annual path) and the movement of the nodes of the Moon's orbit, all the stars in a belt of 10° around the ecliptic are occulted at some time during a period of about nine years. Among these are the bright stars Aldebaran, Regulus, Spica, and Antares, the star clusters Pleiades, Hyades, and Praesepe, and the Crab Nebula.

Since the Moon always moves eastward, an occulted star disappears at the Moon's eastern limb and reappears at the western. These phenomena can be best observed at the dark limb of the Moon.

Efforts have been made to determine the apparent diameter of stars by accurately estimating the time they take to disappear behind the Moon; with modern optics and electronic devices this would not be impossible in some cases. But there are complications of two kinds: first, many stars are close binaries (doubles); and, second, the decline of the light of the occulted star shows an optical phenomenon called a diffraction pattern. However, for the purpose of yielding basic data on the perturbations (disturbances) in the Moon's orbit and on irregularities in the rotation of the Earth, the exact times of star occultations are still of value.

Griffith Observatory, Paul Roques

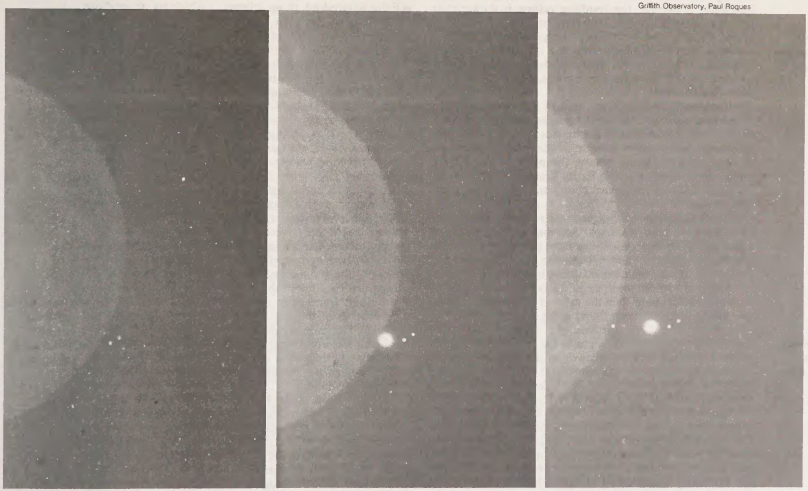


Figure 7: (Left to right) The emergence of Jupiter and three of its satellites after their occultation by the Moon.

The disappearance or reappearance of Jupiter takes more than one minute, but the time for a minor planet is under one second, and that for even a giant star like Aldebaran or Antares is less still, by a factor of 10. Other stars disappear nearly instantaneously, indicating that the Moon has no sensible permanent atmosphere.

ECLIPSING BINARY STARS

When, in a binary system (a double star whose components orbit a common centre), one component comes between the other and the Earth, an eclipse occurs and the amount of light received from both stars together is reduced; a total eclipse gives the greatest reduction, and the light is reduced to a minimum. Two eclipses occur during each revolution. The first star shown to have light that varied through eclipses in this way was Algol (Beta Persei). Its variability was discovered in 1667 by the Italian astronomer Geminiano Montanari, but its periodicity was not discovered and the explanation of the phenomenon as an eclipsing variable was not proposed until a century later by the English astronomer John Goodricke. At present, about 3,000 eclipsing binaries are known.

The many possible sizes of the two stars, in relation both to each other and to the size and tilt of their orbits, combined with the eccentricity (deviation from the circular) of the orbits, lead to a large variety of ways in which the light of the system changes during a revolution: the so-called light curve.

If the orbits of the two stars are not circular, then the motion in the orbits is not uniform but is most rapid near periastron (stars closest together) and least rapid near apastron (stars farthest apart). Therefore, the eclipses may not be equally spaced in time, as they are in the case of circular orbits. Study of the spacing and durations of the light minima will yield both the eccentricity and the orientation of the orbits. A steady change in the spacing of the minima as well as in their relative duration means that the line of apsides (major axis of the relative orbit) is rotating; the most rapid such rotation known is that of the eclipsing binary GL Carinae.

If the two stars are unequal in size, they may wholly overlap for some time. When the star of higher surface brightness is behind, the resulting eclipse is darker than that resulting when the brighter star cuts off the light of the dimmer one. Limb darkening, the tendency for a star to appear less brilliant near the edge of its apparent disk, also influences the shape of the light curves, and the amount of limb darkening present may be determined from those curves.

Limb darkening

Spectroscopic studies of the light of a double star allow determination of the ratio of the masses of the components to each other. The sum of the masses, however, can only be determined if the inclination of the orbit is known. In the case of an eclipsing binary, the plane of the orbit is known to be directed to the Earth. Therefore, eclipsing binaries can be made to reveal complete information about their masses.

A rapidly rotating star bulges at its equator. Also, many double stars are so close together that they distort each other tidally; in the case of very close pairs, this distortion has caused them to turn always the same sides toward each other. If such stars are eclipsing binaries, the amount of distortion can, in some cases, be determined from the light curve. For some stars the ratio of the long axis of the star to the short has been found to be as great as 5:3.

The star Zeta Aurigae

There are a few star systems like Algol, notably Zeta Aurigae, in which the difference in the sizes of the components is so great that the crossing of the smaller, brighter star before the disk of the supergiant companion should really be called a transit rather than an eclipse. This particular star has the remarkably long orbital period of 972 days and consists of an orange supergiant having a diameter of almost 300,000,000 kilometres and a substantially smaller, brighter star, only about three times larger than the Sun. The secondary minimum lasts for 38 days, and a very sensitive photometer must be used to detect the extremely slight fading of light when the smaller star is shining through the extended, tenuous atmosphere of the large one.

ECLIPSES IN HISTORY

In ancient and medieval times, eclipses of both the Sun and the Moon were often regarded as portents; hence, it is not surprising that many of these events are mentioned in history and in literature, as well as in astronomical writings.

Well over 1,000 individual eclipse records are extant from various parts of the ancient and medieval world. Most known ancient observations of these phenomena originate from three countries: Babylonia, China, and Greece. No records appear to have survived from ancient Egypt or India. Whereas virtually all Babylonian accounts of eclipses are confined to astronomical treatises, those from China and Greece are found in historical and literary works as well. Eclipses are noted from time to time in surviving European writings from the early Middle Ages. At this time only the Chinese, however, continued to observe and record such events on a regular basis, and this tradition continued almost uninterrupted down to recent centuries. Many eclipses were carefully observed by the astronomers of Baghdad and Cairo between about AD 800 and 1000. About AD 800 both European and Arab annalists began to include in their chronicles accounts of eclipses and other remarkable celestial phenomena. Some of these chronicles continued until the 14th or 15th century. Toward the end of this period, European astronomers commenced making fairly accurate measurements of the time of day or night when eclipses occurred, and this pursuit spread rapidly following the invention of the telescope.

The value of ancient and medieval records may be classified as follows, although it should be emphasized that there is some overlap between these individual categories: (1) literary and historical, depending on the interest that these records aroused and their connection with historical events; (2) chronological, insofar as they make it possible to verify chronological systems resting on other evidence and to supply dates for events concerned with eclipses; and (3) astronomical, including the determination by ancient astronomers of the periods and motions of the Sun and the Moon and by modern astronomers of variations in the length of the mean solar day.

The Sun is normally so brilliant that the casual observer is liable to overlook those eclipses in which less than about 80 percent of the solar disk is obscured. Only when a substantial proportion of the Sun is covered does the loss of daylight become noticeable. Hence it is rare to find references to small partial eclipses in literary and historical works. At various times, astronomers in Babylonia, China, and the Arab lands systematically reported eclipses of small magnitude but their vigilance was assisted by their ability to make approximate predictions. They thus knew roughly when to scrutinize the Sun. Arab astronomers sometimes viewed the Sun by reflection in water to diminish its brightness when watching for eclipses. The Roman philosopher and writer Seneca (*c.* 4 BC–AD 65), on the other hand, recounts that in his time pitch was employed for this purpose. It is not known, however, whether such artificial aids were used regularly.

When the Moon covers a large proportion of the Sun, the sky becomes appreciably darker and stars may appear. On those rare occasions when the whole of the Sun is obscured, the sudden occurrence of intense darkness, accompanied by a noticeable fall in temperature, may leave a profound impression on eyewitnesses. Total or near-total eclipses of the Sun are of special chronological importance. On average, they occur so infrequently at any particular location that if the date of such an event can be established by historical means to within about a decade, it may well prove possible to fix an exact date by astronomical calculation.

Importance of total or near-total solar eclipses

The Full Moon is much dimmer than the Sun, and lunar eclipses of even quite small magnitude are thus fairly readily visible to the unaided eye. Both partial and total obscurations are recorded in history with comparable frequency. As total eclipses of the Moon occur rather often (every three or four years on average at a given place), they are of less chronological importance than their solar counterparts. There are, however, several notable exceptions to this rule, as will be discussed below.

Literary and historical references. *Old Babylonian.* The earliest known references to eclipses for which dates can be established with reasonable confidence go back to the 21st century BC. These are recorded on the series of astrological tablets from Ur known as *Enūma Anu Enlil.* Several of these texts contain lunar eclipse *omina*—warnings of disasters that might follow an eclipse based on past coincidences between celestial and terrestrial occurrences. Some of the *omina* are so detailed that they are clearly based on observation of a specific eclipse. The example cited below is found on tablet 20 of the series:

> If in Simanu [lunar month III] an eclipse occurs on day 14, the [Moon-] god in his eclipse is obscured on the east side above and clears on the west side below, the north wind blows, [the eclipse] commences in the first watch of the night and it touches the middle watch. . . . by this the [Moon-] god gives a decision for Ur and the king of Ur. The king of Ur will see a famine, there will be many deaths, the king of Ur will be wronged by his son; the son who has wronged his father, the Sun-god will catch him, and he will die at the burial of his father. A son of the king who was not named for kingship will then occupy the throne.

From a careful investigation of the historical and astronomical circumstances, it has been shown that the eclipse referred to here is very likely to have been associated with the murder of Shulgi by his son and the accession of Amar-Sin. The most probable date for the eclipse is April 4, 2094 BC. A further lunar eclipse 42 years later regarded as signaling the destruction of Ur has been dated to April 13, 2053 BC.

Chinese. According to long-established tradition, the history of astronomy in ancient China could be traced back to before 2000 BC. The earliest surviving relicts that are of astronomical significance date from nearly a millennium later, however. The An-yang oracle bones (inscribed turtle shells, ox bones, and so forth) of the Shang dynasty (*c.* 1550–1050 BC), which have been uncovered near An-yang in northeastern China, record several eclipses of both the Sun and the Moon. The following report is an example:

Early Chinese references

> On day *kuei-yu* [the 10th day of a 60-day cycle], it was inquired [by divination]: "The Sun was eclipsed in the evening; is it good?" On day *kuei-yu* it was inquired: "The Sun was eclipsed in the evening; is it bad?"

The above text provides clear evidence that eclipses were regarded as omens at this early period (as is true of other celestial phenomena). Such a belief was extremely prevalent in China during later centuries. The term translated here as "eclipse" (*chih*) is the same as the word "eat." The Shang people thought that some monster was actually devouring the Sun or Moon during an eclipse. Not until many centuries later was the true explanation known; but by then the use of the term *chih* was firmly established to describe eclipses, and so it continued throughout Chinese history. As the year in which an eclipse occurred is never mentioned on the preserved oracle bones (many of which are mere fragments), dating of these observations by astronomical calculation has proved extremely difficult. A recent investigation of five lunar eclipses yielded likely dates between 1200 and 1180 BC. Shang chronology, however, is still very uncertain.

The *Shih ching* ("Classic of Poetry") contains a lamentation occasioned by an eclipse of the Moon followed by an eclipse of the Sun. The text, dating from the 8th century BC, may be translated:

> The Sun was eclipsed, a thing of very evil omen. Then the Moon became small, and now the Sun became small. . . . For the Moon to be eclipsed is but an ordinary matter. Now that the Sun has been eclipsed—how bad it is!

The solar eclipse is said to have occurred on the day *hsin-mao* (the 28th day of the sexagenary cycle), which was the first day of the 10th lunar month. A date of 776 BC was formerly adopted for such an event, but modern computations show that no solar eclipse in that year was visible in China. A revised date of 735 BC has been proposed. The different attitudes toward solar and lunar eclipses at this time is interesting. Throughout the subsequent thousand years or so, lunar eclipses were hardly ever reported in China—in marked contrast to solar obscurations, which were systematically observed.

After about 200 BC, a wide variety of celestial phenomena began to be noted in China on a regular basis. Summaries of these records are found in astronomical treatises contained in the official dynastic histories. In many instances, a report is accompanied by a detailed astrological prognostication. For example, the *Hou-Han shu* ("History of the Later Han Dynasty") contains the following account under a year corresponding to AD 119–120.

> On the day *wu-wu*, the 1st day of the 12th lunar month, the Sun was eclipsed; it was almost complete. On the Earth it became like evening. It was 11 degrees in the constellation of the Maid. The woman ruler [*i.e.*, the Empress Dowager] showed aversion to it. Two years and three months later, Teng, the Empress Dowager, died.

The date of this eclipse on the Chinese calendar is equivalent to January 18, AD 120. On this exact day there had occurred an eclipse of the Sun that was very large in China. Such chronological precision is typical of almost all Chinese records of celestial phenomena. The above-cited text is particularly interesting because it clearly describes an obscuration of the Sun, which, though causing dusk conditions, was not quite total where it was seen. The place of observation was probably Lo-yang, the Chinese capital of the time. With regard to the accompanying prognostication, it should be pointed out that a delay of two or three years between the occurrence of a celestial omen and its presumed fulfillment is quite typical of Chinese astrology.

Assyrian. The Assyrian eponym canon, which preserves the names of the annual magistrates who gave their names to the years (similar to the Athenian archons or Roman consuls), records under the year that corresponds to 763–762 BC: "Insurrection in the city of Ashur. In the month Sivan [equivalent to May–June], the Sun was eclipsed." The reference must be to the eclipse of June 15, 763 BC, the only large solar eclipse visible in Assyria over a period of many years. A possible allusion to the same eclipse is found in the Old Testament: " 'And on that day,' says the Lord God, 'I will make the sun go down at noon, and darken the earth in broad daylight' " (Amos 8:9). Amos was prophesying during the reign of King Jeroboam II (786–746 BC), and the eclipse would be very large throughout Israel.

Jewish. Apart from Amos, the only Old Testament writer to allude to eclipses is Joel (2:31). The most direct account of an eclipse in ancient Jewish history occurs not in the Bible but in the writings of Flavius Josephus, the 1st-century-AD historian. Not long before the death of Herod the Great, Josephus recounts the occurrence of a lunar obscuration:

References in early Jewish writings

> As for the other Matthias who had stirred up the sedition, he [Herod] had him burned alive along with some of his companions. And on that same night there was an eclipse of the Moon. But Herod's illness became more and more severe. . . .

This eclipse occurred shortly before the Passover festival. Calculation shows that the only springtime lunar eclipses visible in Israel between 17 BC and AD 3 took place on March 23, 5 BC and March 13, 4 BC. The former was total, while on the latter occasion about one-third of the Moon was covered. These two dates are conveniently close to one another, although the latter date is usually preferred by chronologists, implying that Herod died in the spring of 4 BC.

Greek. In a fragment of a lost poem by Archilochus occur the words:

> Nothing there is beyond hope, nothing that can be sworn impossible, nothing wonderful, since Zeus, father of the Olympians, made night from mid-day, hiding the light of the shining Sun, and sore fear came upon men.

This seems a clear reference to a total eclipse. The phenomenon has been identified as the eclipse on April 6, 648 BC, which was total in the Aegean and occurred during Archilochus' lifetime.

Small fragments survive of other early Greek poetic descriptions of eclipses, and the ninth paean of Pindar, addressed to the Thebans, takes an eclipse of the Sun as its theme, as follows:

> Beam of the Sun! O thou that seest from afar, what wilt thou be devising? O mother of mine eyes! O star supreme,

reft from us in the daytime! Why hast thou perplexed the power of man and the way of wisdom, by rushing forth on a darksome track?

Pindar then proceeds to speculate on the meaning of this omen. Although he prays, "Change this worldwide portent into some painless blessing for Thebes," he adds, "I in no wise lament whate'er I shall suffer with the rest." This strongly suggests that Pindar, who was a Theban, had himself recently witnessed a great eclipse at his hometown. The most probable date for the eclipse is April 30, 463 BC; modern calculations indicate that the eclipse was nearly total at Thebes.

The historian Thucydides comments on the frequency of eclipses during the Peloponnesian War, which began in 431 BC and lasted for 27 years. The most interesting of these was a solar eclipse that occurred in the summer of the first year of the war (calculated date August 3, 431 BC) and a lunar eclipse that took place in the summer of the 19th year (calculated date August 27, 413 BC). On the former occasion, "the Sun assumed the shape of a crescent and became full again, and during the eclipse some stars became visible" (a statement that agrees well with modern computations). The latter date had been selected by the Athenian commanders Nicias and Demosthenes for the departure of their armies from Syracuse. All preparations were ready, but the signal had not been given when the Moon was eclipsed. The Athenian soldiers and sailors clamoured against departure, and Nicias, in obedience to the soothsayers, resolved to remain thrice nine days. This delay enabled the Syracusans to capture or destroy the whole of the Athenian fleet and army.

August 15, 310 BC, is the date of a total eclipse of the Sun that is said to have been seen at sea by Agathocles and his men after they had escaped from Syracuse and were on their way to Africa. Diodorus, a historian of the 1st century BC, reports that, "On the next day [after the escape] there occurred such an eclipse of the Sun that utter darkness set in and the stars were seen everywhere." Modern computations of the eclipse track render it probable that Agathocles' ships passed along the north of Sicily during the course of the journey; the Sun would have only been partially obscured on the south side of the island.

Plutarch's interest in eclipses

In Plutarch's dialogue concerning the features of the Moon's disk, one of the characters, named Lucius, deduces from the phases of the Moon and the phenomenon of eclipses a similarity between the Earth and the Moon and illustrates his argument by means of a recent eclipse of the Sun, "which, beginning just after noon, showed us plainly many stars in all parts of the heavens, and produced a chill in the temperature like that of twilight." This eclipse has been identified with one that occurred on March 20, AD 71, which was total in Greece. Whether Plutarch is describing a real, and therefore datable, event or is merely basing his description on accounts written by earlier authors has been disputed, however. Later in the same dialogue, Lucius refers to a brightness that appears around the Moon's rim in total eclipses of the Sun. This is one of the earliest known allusions to the solar corona. Plutarch was unusually interested in eclipses, and his *Parallel Lives*, an account of the deeds and character of illustrious Greeks and Romans, contains many references to both lunar and solar eclipses of considerable historical importance. There also are frequent records of eclipses in other ancient Greek literature.

Roman. Roman history is less replete with references to eclipses than that of Greece, but there are several interesting references to these events in Roman writings. Some, like the total solar eclipse said by Dio Cassius, a Roman historian of the 3rd century AD, to have occurred at the time of the funeral of Agrippina, the mother of Nero, never took place. One that has attracted the attention of students of astronomy and of the Roman calendar alike is stated by Cicero to have occurred in what may have been the 350th year from the founding of Rome. He also says that it was described by the poet Quintus Ennius: "On the nones of June the Sun was covered by the Moon and night." This happening would appear to have been the total solar eclipse of June 21, 400 BC, which reached a total or almost total phase at Rome a few minutes after sunset. Its recorded date seems to show that in that year the calendar month of June began 16 days later than it did after the Julian reform. The eclipse of the Moon on June 21–22, 168 BC, has attracted much attention. The Romans were at that time at war with Macedonia, and Polybius says that this eclipse was interpreted as an omen of the eclipse of a king and thus encouraged the Romans and discouraged the Macedonians.

What may well be an indirect allusion to a total eclipse of the Sun that caused darkness at Rome is recorded by Livy for a time corresponding to 188–187 BC (the consulship of Valerius Messalla and Livius Salinator):

> Before the new magistrates departed for their provinces, a three-day period of prayer was proclaimed in the name of the College of Decemvirs at all the street-corner shrines because in the daytime at the third hour darkness had covered everything.

The darkness took place some time after the election of the consuls (Ides of March), and, allowing for the confusion of the Roman calendar at this time, the total eclipse of July 17, 188 BC, would be the most satisfactory explanation for the unusual darkness. Since the Sun is not mentioned in the text, the phenomenon possibly occurred on a cloudy day. Two years earlier (190 BC), Livy records an eclipse as happening at the beginning of July. The calculated date, however, is March 14 in that year. Consequently, the Roman calendar in that year must have been as much as 3½ months out of adjustment.

Medieval European. Following the close of the Classical Age, eclipses were in general only rarely recorded by European writers for several centuries. Not until after about AD 800 did eclipses and other celestial phenomena begin to be frequently reported again, especially in monastic chronicles. Hydatius, bishop of Chaves (in Portugal), was one of the few known chroniclers of the early Middle Ages. He seems to have had an unusual interest in eclipses, and he recounted the occurrence of five such events (involving both the Sun and the Moon) between AD 447 and 464. In each case, only brief details are given, and Hydatius gives the years of occurrence in terms of the Olympiads (*i.e.*, reckoning time from the first Olympic Games, in 776 BC). During the total lunar eclipse of March 2, AD 462 (this date is known to be accurate), the Moon is said to have been "turned into blood." Statements of this kind are common throughout the Middle Ages, presumably inspired by the Old Testament allusion in Joel (2:31). Similar descriptions, however, are occasionally found in non-Christian sources, as, for example, a Chinese one of AD 498.

Given below is a selection from the vast number of extant medieval European reports of eclipses. In many cases, the date is accurately recorded, but there also are frequent instances of chronological error.

English

An occultation of a bright star by the eclipsed Moon in AD 756 (actually the previous year) is the subject of an entry in the chronicle of Simeon of Durham, compiled some four centuries after the event:

> Moreover, the Moon was covered with a blood-red color on the 8th day before the Kalends of December [*i.e.*, November 24] when 15 days old, that is, the Full Moon; and then the darkness gradually decreased and it returned to its original brightness. And remarkably indeed, a bright star following the Moon itself passed through it, and after the return to brightness it preceded the Moon by the same distance as it had followed the Moon before it was obscured.

The text gives no hint of the identity of the star. Modern computations show that the Moon was totally eclipsed on the evening of November 23, AD 755. During the closing stages of the eclipse, Jupiter would have been occulted by the Moon, as seen from England. This is an example of the care with which an observer who was not an astronomer could describe a compound astronomical event without having any real understanding of what was happening.

Byzantine

Several eclipses are recorded in Byzantine history, beginning in the 6th century AD. By far the most vivid account relates to the solar eclipse of December 22, AD 968. This was penned by the contemporary chronicler Leo the Deacon:

> At the winter solstice there was an eclipse of the Sun such as has never happened before.... It occurred on the 22nd

day of the month of December, at the 4th hour of the day, the air being calm. Darkness fell upon the Earth and all the brighter stars revealed themselves. Everyone could see the disk of the Sun without brightness, deprived of light, and a certain dull and feeble glow, like a narrow headband, shining round the extreme parts of the edge of the disk. However, the Sun gradually going past the Moon (for this appeared covering it directly) sent out its original rays, and light filled the Earth again.

This is the earliest account of the solar corona that can be definitely linked to a datable eclipse. Although the appearance of the corona during totality is rather impressive, early descriptions of it are extremely rare. Possibly many ancient and medieval eyewitnesses were so terrified by the onset of sudden darkness that they failed to notice that the darkened Sun was surrounded by a diffuse envelope of light.

Italian

In a chronicle of the Norman rule in Sicily and southern Italy during the 11th century, Goffredo Malaterra records an eclipse of the Sun that, even though it caused alarm to some people, was evidently regarded by others as no more than a practical inconvenience:

[AD 1084] On the sixth day of the month of February between the sixth and ninth hours the Sun was obscured for the space of three hours; it was so great that any people who were working indoors could only continue if in the meantime they lit lamps. Indeed some people went from house to house to get lanterns or torches. Many were terrified.

This eclipse actually occurred on February 16, AD 1086. It was the only large eclipse visible in southern Italy for several years around this time; hence, the chronicler had mistaken both the year and day.

Medieval Islāmic. Like their Christian counterparts, medieval Islāmic chroniclers record a number of detailed and often vivid descriptions of eclipses. Usually the exact date of occurrence is given (on the lunar calendar). A graphic narrative of the total solar eclipse of June 20, AD 1061, is recorded by the Baghdad annalist Ibn al-Jawzī, who wrote approximately a century after the event:

On Wednesday, when two nights remained to the completion of the month Jumādā I [in AH 453], two hours after daybreak, the Sun was eclipsed totally. There was darkness and the birds fell whilst flying. The astrologers claimed that one-sixth of the Sun should have remained [uneclipsed] but nothing of it did so. The Sun reappeared after four hours and a fraction. The eclipse was not in the whole of the Sun in places other than Baghdad and its provinces.

The date corresponds exactly to June 20, AD 1061, on the morning of which there was a total eclipse of the Sun visible in Baghdad. The duration of totality is much exaggerated, but this is common in medieval accounts of eclipses. The phenomenon of birds falling from the sky at the onset of the total phase was also noticed in Europe during several eclipses in the Middle Ages.

References in Islāmic texts

Two independent accounts of the total solar eclipse of AD 1176 are recorded in contemporary Arab history. Ibn al-Athīr, who was age 16 at the time, described the event as follows:

In this year [AH 571] the Sun was eclipsed totally and the Earth was in darkness so that it was like a dark night and the stars appeared. That was the forenoon of Friday the 29th of the month Ramaḍān at Jaẓīrat Ibn 'Umar, when I was young and in the company of my arithmetic teacher. When I saw it I was very much afraid; I held on to him and my heart was strengthened. My teacher was learned about the stars and told me, "Now, you will see that all of this will go away," and it went quickly.

The date of the eclipse is given correctly apart from the weekday (actually Sunday) and is equivalent to April 11, AD 1176. Calculation shows that the whole of the Sun would have been obscured over a wide region around Jaẓīrat Ibn 'Umar (now Cizre in Turkey). Farther south, totality was also witnessed by Saladin and his army while crossing the Orontes River near Ḥamāh (in present-day Syria). The chronicler 'Imād al-Dīn, who was with Saladin at the time, noted that, "The Sun was eclipsed and it became dark in the daytime. People were frightened and stars appeared." As it happens, 'Imād al-Dīn dates the event one year too early (AH 570), but the only large eclipse visible in this region for several years occurred in AD 1176.

Lunar and solar eclipses are fairly frequently visible on the Earth's surface 15 days apart, and from time to time such a pair of eclipses may be seen from one and the same location. Such was the case in the summer of AD 1433, but this occurrence caused some surprise to the contemporary Cairo chronicler al-Maqrīzī:

On Wednesday the 28th of Shawwāl [*i.e.*, June 17], the Sun was eclipsed by about two-thirds in the sign of Cancer more than one hour after the afternoon prayer. The eclipse cleared at sunset. During the eclipse there was darkness and some stars appeared. . . . On Friday night the 14th of Dhu l-Qu'da [July 3], most of the Moon was eclipsed. It rose eclipsed from the eastern horizon. The eclipse cleared in the time of the nightfall prayer. This is a rarity—the occurrence of a lunar eclipse 15 days after a solar eclipse.

The loss of daylight produced by the solar eclipse is much exaggerated but otherwise the description is fairly careful.

Uses of eclipses for chronological purposes. Several examples of the value of eclipses in chronology have already been mentioned in passing. No one system of dating has been continuously in use since ancient times, although some, like the Olympiads, persisted for many centuries. Dates were frequently expressed in terms of a king's reign; years were also named after officials of whom lists have been preserved (the eponym canons mentioned above). In such cases, it is important to be able to equate certain specific years thus defined with years before the Christian Era (BC). This correspondence can be made whenever the date of an eclipse is given in an ancient record. In this regard, eclipses have distinct advantages over other celestial phenomena such as comets: in addition to being frequently recorded in history, their dates of occurrence can be calculated exactly.

Chinese chronology can be confirmed accurately by eclipses from the 8th century BC (during the Chou dynasty) onward. The *Ch'un-ch'iu* ("Spring and Autumn Annals"), a chronicle covering the period from 722 to 481 BC, notes the occurrence of 36 solar eclipses during this interval. This is the earliest surviving series of eclipse observations from any part of the world. The records give the date of each event in the following form: year of the ruler, lunar month, and day of the 60-day cycle. Three of the eclipses (occurring in 709, 601, and 549 BC) were described as total. As many as 32 of the eclipses cited in the *Ch'un-ch'iu* can be identified by modern calculations. Errors in the recorded lunar month (typically amounting to no more then a single month) are fairly common, but both the year and the recorded day of the sexagenary cycle are invariably correct.

Earliest surviving series of eclipse observations

The chronology of Ptolemy's canon list of kings, which gives the Babylonian series from 747 to 539 BC, the Persian series from 538 to 324 BC, the Alexandrian series from 323 to 30 BC, and the Roman series from 30 BC onward, is confirmed by eclipses. The eclipse of 763 BC, recorded in the Assyrian eponym canon, makes it possible to carry the chronology back with certainty through the period covered by that canon to 893 BC. Identifiable eclipses that were recorded under named Roman consuls extend back to 217 BC. The dated solar eclipse of Ennius (400 BC), the lunar eclipse seen at Pydna in Macedonia on June 21–22, 168 BC, and the solar eclipse recorded at Rome in 190 BC can be used to determine months in the Roman calendar in the natural year. Furthermore, eclipses occasionally help to fix the precise dates of a series of events, such as those associated with the Athenian disaster at Syracuse.

The late Babylonian astronomical texts occasionally mention major historical events, as, for example, the dates when Xerxes and Alexander the Great died. Most of these clay tablets, inscribed with a cuneiform script, are now found in the British Museum. The preserved texts are mainly in the form of day-to-day diaries of celestial observations and summary tables that abstract specific types of observation from the diaries. To illustrate the potential of this material for chronological purposes, the date of the death of Xerxes may be accurately fixed by reference to eclipses. On a tablet that lists lunar eclipses at 18-year intervals occurs the following brief announcement between

two eclipse records: "Month V, day 14 [?], Xerxes was murdered by his son." Unfortunately, the cuneiform sign for the day of the month is damaged, and a viable reading could be anything from 14 to 18. The year is missing, but it can be deduced from the 18-year sequence as 465 BC. This identification is confirmed by calculating the dates of the two eclipses stated to have occurred in the same year that Xerxes died. The first of these happened when the Moon was in the constellation of Sagittarius, while the second took place on the 14th day of the 8th lunar month. For many years both before and after 465 BC, no such combination of eclipses can be found; it occurs only in 465 BC itself. The dates deduced for the two eclipses are June 5 and November 30 of that year. Mention of an intercalary sixth month on the same tablet enables the date of the death of Xerxes to be fixed as some time between August 4 and 8 in 465 BC.

Uses of eclipses for astronomical purposes. *In ancient astronomy.* It is known from both Babylonian and Greek history that at least from the time of King Nabonassar (whose reign began in 747 BC), a dated canon of astronomical observations was preserved at Babylon. This included numerous eclipses of both the Sun and the Moon. Reference to the list of lunar eclipses would enable the Babylonian astronomers to determine accurately the intervals between such eclipses and must have facilitated the discovery of the 18-year cycle (more exactly the cycle of 6,585⅓ days that the 10th-century Greek lexicographer Suidas named the saros). This cycle is attested in Babylonian astronomy at a fairly early period. In contrast to lunar obscurations, the visibility of solar eclipses at a given place on the Earth's surface is complicated by geographic considerations. Hence, even a lengthy list of solar eclipse dates would be of limited value in deducing the saros.

Babylonian list of lunar eclipses

The surviving late Babylonian astronomical texts contain many examples of predictions of both lunar and solar eclipses using cycles. Comparison with modern computations shows that, although some predictions were successful, in other instances an eclipse was visible only elsewhere on the Earth's surface or did not occur at all. A theory based on a detailed knowledge of apparent lunar and solar motions is necessary to enable eclipses to be predicted accurately. Lunar eclipses indicate more accurately than any other phenomena the actual time when the Sun and the Moon are in opposition. From an early date the Babylonian astronomers must have deduced from lunar eclipses not only the mean interval between two conjunctions (closest apparent approaches) of the Sun and the Moon but also the principal inequality (change of speed) in the motion of the Moon and the similar inequality in the motion of the Earth (or, as according to their geocentric theory they conceived it, of the Sun). The Babylonians were able to define the periods of these inequalities (the cause of which lies in the ellipticities of the orbits), which astronomers refer to as the anomalistic month and year.

In the same way, since eclipses happen only when the Sun and the Moon are at the intersections of their orbital planes called nodes, and since the path of the shadow in a lunar eclipse depends on the position of the centre of the Sun in relation to the node, the Babylonians were also able to determine the position and motion of the nodes. By assuming, as is approximately true, that the saros of 6,585⅓ days contained an exact number: (1) of synodic months, or revolutions of the Moon measured from the Sun, (2) of anomalistic months, or revolutions of the Moon measured from its apogee or perigee (*i.e.*, from its farthest distance from and closest approach to the Earth), and (3) of draconic months, or revolutions of the Moon measured from its node, astronomers, perhaps as early as the 6th century BC, computed the relative motions of the Sun and the Moon, the lunar perigee and apogee, and the nodes. About 500 BC the Babylonian astronomer Nabu-rimmani, apparently from a more accurate study of eclipse observations, obtained improved values that: for the motion of the Moon relative to the Sun were 10″ of arc per annum too small; for the Moon's perigee motion 20″ of arc per annum too great; and for the motion of its node 5″ of arc too small. Still more accurate values were obtained by Kidinnu about 383 BC, from whom they passed to the Greek astronomer Hipparchus. In the system of Nabu-rimmani the distance of the Moon from its node was used for the prediction of the magnitude of lunar eclipses.

In modern astronomy. Ancient and medieval observations of eclipses are of the highest value for investigating long-term variations in the length of the day. Early investigators such as Edmond Halley deduced from eclipse observations that the Moon's motion was subject to an acceleration. However, not until 1939 was it demonstrated (by Harold Spencer Jones) that only part of this acceleration was real. The remainder was apparent and was a consequence of the practice of measuring time relative to a non-uniform unit—namely, the rotation of the Earth. Time determined in this way is termed Universal Time. For astronomical purposes, it is preferable to utilize an invariant time-frame such as Ephemeris Time.

Lunar and solar tidal friction, occurring especially in the seas and oceans of the Earth, is now known to be responsible for a gradual decrease in the terrestrial rate of rotation. Apart from slowing down the Earth's rotation, lunar tides produce a reciprocal effect on the Moon's motion, causing a gradual increase in the mean distance of the Moon from the Earth (at about 3.5 centimetres [1.4 inches] per year) and a consequent real retardation of its motion. Hence, the length of the month is slowly increasing. These changes in the Moon's orbit can now be accurately fixed by lunar laser ranging, and it seems likely that they have proceeded at an essentially constant rate for many centuries. The history of the Earth's rotation, however, is complicated by effects of nontidal origin, and in order to obtain maximum information it is necessary to utilize both modern and ancient observations. Telescopic observations reveal fluctuations in the length of the day on time scales as long as decades, and these fluctuations are mainly attributed to interactions between the fluid core of the Earth and the surrounding solid mantle. Ancient and medieval observations also suggest the presence of longer term variations, which could be produced by alterations in the moment of inertia of the Earth resulting from both the ongoing rise of land that was glaciated during the Pleistocene ice age and sea-level changes associated with the freezing and melting of polar ice.

Studying past variations in the Earth's rotation rate

Records of large solar eclipses preserved in literary and historical works have made an important contribution to the study of past variations in the Earth's rate of rotation. Nevertheless, the major contribution has come from the analysis of timings of lunar and solar eclipses by ancient Babylonian and medieval Arab astronomers. (Unfortunately, there are very few measurements of intermediate date.) Although many Babylonian texts are fragmentary, nearly 100 usable timings of eclipse contacts are accessible (including measurements at different phases of the same eclipse). These observations date primarily from between about 550 and 50 BC. By comparison, only a handful of similar Greek measurements are preserved, and these are far less precise. About 50 eclipse timings by medieval Arab astronomers are preserved; these are mainly contained in the Hakemite Tables compiled by Ibn Yūnus about AD 1005.

Tidal computations indicate a steady increase in the length of the mean solar day by about 1/40 second every millennium, with other nontidal causes producing additional smaller effects. Although this seemingly represents a minute change, the long time scale covered by ancient observations is an important asset. Approximately one million days—each marginally shorter than at present—have elapsed since the earliest reliable eclipse observations were made, about 700 BC. As a result, present-day computations of ancient eclipses that make no allowance for any increase in the length of the day may be as much as five or six hours ahead of the observed time of occurrence. In the case of total solar eclipses, the path of the Moon's shadow across the Earth's surface may appear to be displaced by thousands of kilometres.

The technique of using ancient observations to investigate changes in the rate of the Earth's rotation is well illustrated by a total solar eclipse observed by Babylonian astronomers on a date corresponding to April 15 in 136

BC. This event is recorded on two damaged tablets, a composite translation of which follows:

> At 24 degrees after sunrise, there was a solar eclipse beginning on the southwest side. After 18 degrees it became total such that there was complete night. Venus, Mercury, and the normal stars were visible. Jupiter and Mars, which were in their period of disappearance, were visible in that eclipse. [The shadow] moved from southwest to northeast. [Time interval of] 35 degrees for obscuration and clearing up.

This is an exceptionally fine account of a total solar eclipse and is by far the best preserved from the ancient world. As will be seen, the Babylonians were able to detect a number of stars, as well as four planets, during the few minutes of darkness. Modern calculations confirm that Jupiter and Mars were too near the Sun to be observed under normal circumstances; Jupiter was very close to the solar disk.

Time intervals were expressed by the Babylonians in degrees, each equivalent to 4 minutes of time. Hence the eclipse is recorded as beginning 96 minutes after sunrise (or about 7:10 AM), becoming total 72 minutes later and lasting from start to finish for 140 minutes. Computations that make no allowance for changes in the length of the day suggest that this eclipse was barely visible at Babylon, with as little as 15 percent of the Sun being covered. Furthermore, the computed time of onset is around noon rather than in the early morning. In order to best comply with the record, it is necessary to assume that the length of the day has increased by about $^1/_{20}$ second in the intervening two millennia.

Numerous eclipses of both the Sun and the Moon were timed by the Babylonian astronomers with similar care, and analysis of the available records closely confirms the above result for the change in the length of the day. Presumably some kind of clepsydra, or water clock, was used to measure time intervals. Although such a device is likely to have been of low precision, many eclipse observations were made fairly close to the reference moments of sunrise or sunset. Hence, the measured intervals would be so short that clock errors may be presumed to be small.

The latest known Babylonian observations date from about 50 BC. After this date, eclipse measurements of comparable precision—by Arab astronomers—are not found until AD 800. The following observations of the lunar eclipse of September 17, AD 1019, made by al-Bīrūni at Ghazna (now Ghazni, Afghanistan) attest to the quality of these more recent data:

> When I observed it, the altitude of Capella above the eastern horizon was slightly less than 60 degrees when the cut at the edge of the Full Moon had become visible; the altitude of Sirius was [then] 17 degrees, that of Procyon was 22 degrees and that of Aldebaran was 63 degrees, where all altitudes are measured from the eastern horizon.

All of these various measurements are in agreement that the eclipse began at around 2:15 AM, but calculations that make no allowance for any change in the length of the day indicate a time approximately $^1/_2$ hour later.

Combining the various results obtained from analysis of ancient and medieval data, it is possible to show that over the last 1,000 years the average rate of increase in the length of the day was only about two-thirds of what it was in the previous millennium. This emphasizes the importance of nontidal effects in producing changes in the rate of the Earth's rotation period. In sum, the history of the Earth's rotation is extremely complex. (J.H./F.R.S.)

BIBLIOGRAPHY. BRYAN BREWER, *Eclipse* (1978); and DAVID ALLEN and CAROL ALLEN, *Eclipse* (1987), explain and recount the history of eclipses for the general reader. DONALD H. MENZEL and JAY M. PASACHOFF, "Solar Eclipse: Nature's Super Spectacular," *National Geographic*, 138(2):222–233 (August 1970), chronicles the events of a solar eclipse expedition. FRANK DYSON and R.v.d.R. WOOLLEY, *Eclipses of the Sun and Moon* (1937); and J.B. ZIRKER, *Total Eclipses of the Sun* (1984), discuss in considerable detail the history, methods, and results of eclipse observations. W.M. SMART, *Textbook on Spherical Astronomy*, 6th ed., rev. by R.M. GREEN (1977), presents the basic mathematical tools for calculating occultations and eclipses. GREAT BRITAIN NAUTICAL ALMANAC OFFICE, *Explanatory Supplement to the Astronomical Ephemeris and the American Ephemeris and Nautical Almanac* (1961, reissued 1977), presents in comprehensive fashion data for the calculation of astronomical phenomena. F. LINK, *Eclipse Phenomena in Astronomy* (1969), treats modern developments in eclipse problems, excluding solar eclipses and eclipsing variables. Works on historical eclipses include F. RICHARD STEPHENSON, "Historical Eclipses," *Scientific American*, 247(4):170–183 (October 1982); SAID S. SAID, F. RICHARD STEPHENSON, and WAFIQ RADA, "Records of Solar Eclipses in Arabic Chronicles," *Bulletin of the School of Oriental and African Studies*, 52:38–64 (1989); and F. RICHARD STEPHENSON and S.S. SAID, "Non-tidal Changes in the Earth's Rate of Rotation as Deduced from Medieval Eclipse Observations," *Astronomy and Astrophysics*, 215(1):181–189 (1989). Catalogs of eclipse data include TH. RITTER VON OPPOLZER, *Canon of Eclipses* (1962; originally published in German, 1887), astronomical data of all solar eclipses between 1207 BC and AD 2161 and eclipses of the Moon from 1206 BC to AD 2163, with maps of the central lines of the solar eclipses over the Earth; GEORGE VAN DEN BERGH, *Eclipses in the Second Millennium B.C. (−1600 to −1207)* (1954), a demonstration of a method for computing each of these eclipses with simple arithmetic, and *Periodicity and Variation of Solar (and Lunar) Eclipses* (1955; originally published in Dutch, 1951), an arrangement of all the eclipses in Oppolzer's *Canon* into the saros and the inex periods in a newly developed panorama; ROBERT R. NEWTON, *Ancient Astronomical Observations and the Accelerations of the Earth and Moon* (1970), and *Medieval Chronicles and the Rotation of the Earth* (1972); and HERMANN MUCKE and JEAN MEEUS, *Canon of Solar Eclipses −2003 to +2526* (1983), and *Canon of Lunar Eclipses −2002 to +2526*, 2nd ed. (1983), more accurate data and maps of central lines over the Earth of all solar and lunar eclipses over 4,500 years. JOSEPH NEEDHAM, *Science and Civilisation in China*, vol. 3, *Mathematics and the Sciences of the Heavens and the Earth* (1959), includes a discussion of eclipses placed in the context of Chinese astronomy with extensive references to original literature. BERNARD LOVELL (ed.), *Astronomy*, 2 vol. (1970), contains discourses in the physical sciences from 1851 to 1939, a large number of which are devoted to solar eclipses. (J.H./F.R.S.)

Economic Growth and Planning

Economic growth involves increases over time in the volume of a country's per capita gross national product (GNP) of goods and services. Such continuing increases can raise average living standards substantially and provide a stronger base for other policy objectives such as national defense, various kinds of capital investments, or public welfare services. It is only in the last two centuries that continued growth in living standards has been realized for a number of now-developed countries, and this process has broadened in the 20th century to include a number of developing countries. However, the fairly steady expansion in the third quarter of the 20th century gave way to a period of slower and more erratic growth for both high- and low-income countries, while some of the economically poorest countries were thus far unable to establish a self-sustaining pattern of development. It also became increasingly evident that there were serious environmental problems associated with some types of growth in production.

This article examines the record of economic growth and development, some explanations for the changes involved, and the attempts by governments to plan these changes. Five major issues are involved. The first is why economic growth occurs more quickly in some countries and periods than in others. It is the increase in the size and quality of the factors of production that underlies growth, but certain forces—innovations and entrepreneurship, the part played by governments, and the role of investment as distinct from consumption—deserve special attention. A variety of models of economic growth give expression to the understanding of these forces. Increasing attention has been paid in these models and in policy to the international aspects of growth. This trend is partly a reflection of the growing internationalization of economic activity. It also reflects a number of potentially destabilizing changes in the international economy that became evident during the 1970s. While the precise nature of their effects is open to debate, among these changes should be noted the transition to more flexible exchange rates, the supply shocks in petroleum and other products, the growth of international debt, and the development of several major centres of economic power.

A second issue is the challenges facing the low-income countries, namely, to move from subsistence levels of per capita income to a level that would generate self-sustaining growth and also to reduce the gap between themselves and the higher-income countries. Differences among the lower-income countries warn against making sweeping generalizations on the development process, but three topics have attracted much attention. One is how far existing private and public organizations must be changed so as to institutionalize development. A second is the view that, particularly for manufactures and for smaller markets, reliance on import substitution should give way fairly quickly to export development. The third is the impact of population growth on both development and living standards.

The uneven patterns of growth and development have led to many strains. In the case of higher-income countries these have appeared particularly as declines and failures in some of their older industries in the face of increased competition among themselves and with the newly industrializing countries. In many developing countries there have been repeated calls for a new organization of world institutions geared to a more equitable distribution of wealth.

A third issue, productivity, is central to changes in living standards and to the analysis of international competitiveness. Productivity is the ratio of what is produced to what is required to produce it, a ratio beset with measurement problems. Studies of the reasons for the growth of productivity have focused on technological change and on the accumulation of physical and human capital. However, a considerable number of social, political, and economic issues appear to be involved in the striking differences in rates of change in productivity among countries. Nor is there agreement on the reasons for and significance of the marked slowing of productivity growth in many countries during the 1970s and 1980s. Some economists see merely passing factors at work, while others predict continuing problems for developed countries and the international economy generally.

A fourth major issue is the attempt to maintain growth and increase development through economic planning. Such planning has other objectives as well, such as regional development, constraining wage–price inflation, and easing the adjustment between declining and growing sectors. Planning became a widespread phenomenon during and just after World War II and was given further emphasis in many newly independent countries that were industrializing. The degree of detail and the methods used in planning range from heavily centralized direction and substantial public ownership to attempts simply to trim the overall balance of supply and demand, with different degrees of attention to industrial strategies along the way. Beginning in the 1970s the emphasis shifted to more decentralized planning, with deregulation and privatization of industry as two aspects of this process.

Underlying economic growth and planning is a fifth issue, the attempt to predict economic activity. Modern forecasting involves a variety of computer-based techniques at the level of the firm, the country, and the international economy. The accuracy of forecasting has been reduced by increased uncertainty in the global and national economies since the early 1970s. (A.E.Sa.)

For coverage of related topics in the *Macropædia* and *Micropædia*, see the *Propædia*, section 536.

This article is divided into the following sections:

How economies grow 879
  The analysis of growth 880
    Quality improvements in the inputs
    Entrepreneurship
    The play of influences
    The role of government
  The social cost of growth 881
  Theories of growth 882
    Role of the entrepreneur
    The role of investment
    Demand and supply
    Economic stagnation
    Foreign trade
    Mathematical growth theories
Economic development 884
  Economic development as an objective of policy 884
    Motives for development
    The impact of discontent
  A survey of development theories 885
    The hypothesis of underdevelopment
    Development thought after World War II
    Growth economics and development economics
    The missing-component approach
    Surplus resources and disguised unemployment
    Role of governments and markets
  Lessons from development experience 888
    The importance of agriculture
    The role of exports
    The negative effect of controls
    The importance of appropriate incentives
    The role of the international economy
    Population growth

Development of domestic industry
Developing countries and debt
Development in a broader perspective 890
Economic productivity 891
Uses of productivity measurement 891
Index of growth
Measure of efficiency
Wage and price analysis
Factors that determine productivity levels 892
Measurement of productivity 893
Output
Inputs
Historical trends 893
Early industrialization
The postwar growth surge
Economic planning 895
The nature of economic planning 895
Economic planning in Communist countries 896
Planning in the U.S.S.R.
Planning in other Communist countries
Assessment of Soviet-type planning
Economic planning in non-Communist countries 899
Planning in developed countries
Planning in developing countries
Economic forecasting 903
Types of forecasting 903
Forecasting the GNP and its elements
Forecasting for an industry or firm
Long-term forecasting
Forecasting techniques 905
Information on spending
Selection of turning points
The accuracy of economic forecasts 906
Bibliography 906

How economies grow

Growth can best be described as a process of transformation. Whether one examines an economy that is already modern and industrialized or an economy at an earlier stage of development, one finds that the process of growth is uneven and unbalanced. Economic historians have attempted to develop a theory of stages through which each economy must pass as it grows. Early writers, given to metaphor, often stressed the resemblance between the evolutionary character of economic development and human life—*e.g.*, growth, maturity, and decadence. Later writers, such as the Australian economist Colin Clark, have stressed the dominance of different sectors of an economy at different stages of its development and modernization. For Clark, development is a process of successive domination by primary (agriculture), secondary (manufacturing), and tertiary (trade and service) production. For the American economist W.W. Rostow, growth proceeds from a traditional society to a transitional one (in which the foundations for growth are developed), to the "take-off" society (in which development accelerates), to the mature society. Various theories have been advanced to explain the movement from one stage to the next. Entrepreneurship and investment are the two factors most often singled out as critical.

Stages of growth

Economic growth is usually distinguished from economic development, the latter term being restricted to economies that are close to the subsistence level. The term economic growth is applied to economies already experiencing rising per capita incomes. In Rostow's phraseology economic growth begins somewhere between the stage of take-off and the stage of maturity; or in Clark's terms, between the stage dominated by primary and the stage dominated by secondary production. The most striking aspect in such development is generally the enormous decrease in the proportion of the labour force employed in agriculture. There are other aspects of growth. The decline in agriculture and the rise of industry and services has led to concentration of the population in cities, first in what has come to be described as the "core city" and later in the suburbs. In earlier years public utility investment (including investment in transportation) was more important than manufacturing investment, but in the course of growth this relationship was reversed. There has also been a rise in the importance of durable consumer goods in total output. In the U.S. experience, the rate of growth of capital goods production at first exceeded the rate of growth of total output, but later this too was reversed. Likewise, business construction or plant expenditures loomed large in the earlier period as an object of business investment compared to the recent era. Whether other countries will go through the same experience at similar stages in their growth remains to be seen.

National comparisons

Comparative growth rates for a group of developed countries show how uneven the process of growth can be. Partly this unevenness reflects the extraordinary nature of the 1913–50 period, which included two major wars and a severe and prolonged depression. There are sizable differences, however, in the growth rates of the various countries as between the 1870–1913 and 1950–73 periods and the period since 1973. For the most part, these differences indicate an acceleration in rates of growth from the first to the second period and a marked slowdown in growth rates from the second to the current period. Many writers have attributed this to the more rapid growth of business investment during the middle of the three periods.

The relatively high rates of growth for West Germany, Japan, and Italy in the post-World War II period have stimulated a good deal of discussion. It is often argued that "late starters" can grow faster because they can borrow advanced technology from the early starters. In this way they leapfrog some of the stages of development that the early starters were forced to move through. This argument is nothing more than the assertion that late starters will grow rapidly during the period when they are modernizing. Italy did not succeed in growing rapidly and thereby modernizing until after World War II. Together with Japan and Germany it also experienced a large amount of war damage. This has an effect similar to starting late, since recovery from war entails building a stock of capital that will, other things being equal, embody the most advanced technology and therefore be more productive and allow faster growth. The other part of this argument is the assertion that early starters are actually deterred from introducing on a broad front the new technology they themselves have developed. For example, firms in a country that industrialized early may be inhibited from introducing a more modern and efficient means of transportation on a broad scale because there is no guarantee that other firms handling the ancillary loading and unloading tasks will also modernize to make the change profitable.

Related to this is the problem of whether or not per capita income levels and their rates of growth in developed economies will eventually converge or diverge. For example, as per capita incomes of fast growers like the Italians and Japanese approach those of economies that developed earlier, such as the American and British, will the growth rates of the former slow down? Economists who answer in the affirmative stress the similarities in the changing patterns of demand as per capita income rises. This emphasis in turn implies that there is less and less chance to borrow technology from the industrial leaders as the income levels of the late starters approach those of the more affluent. Moreover, rising per capita incomes in an affluent society usually are accompanied by a shift in demand toward services. Therefore, so this argument goes, differences in income levels and growth rates between countries should eventually narrow because of the low growth in productivity in the service sector. The evidence is inconclusive. On the one hand, growth is a function of something more than the ability to borrow the latest technology; on the other hand, it is not clear that productivity must always grow at a slower rate in the service industries.

A rapidly increasing population is not clearly either an advantage or a disadvantage to economic growth. The American Simon Kuznets and other investigators have found little association between rates of population growth and rates of growth of GNP per capita. Some of the fastest growing economies have been those with stable populations. And in the United States, where the rate of growth of population has shown a downward historical trend, the

rate of growth of GNP per capita has increased over the last century and a half. Another finding by Kuznets is that while GNP per capita in 1960 was substantially higher in the United States than in any European country, there was no significant difference in the per capita growth rates of all these countries over the period 1840 to 1960 as a whole. The conclusion is that the United States started from a higher per capita base; this may have been the result of its superior natural resources, especially its fertile agricultural land.

THE ANALYSIS OF GROWTH

To explain why some countries grow more rapidly than others or why a country may grow more rapidly during one period of history than another, economists have found it convenient to think in terms of a "production function." This is a mathematical way of relating some measure of output, such as GNP, to the inputs required to produce it. For example, it is possible to relate GNP to the size of the labour force measured in man-hours, to capital stock measured in dollars, and to various other inputs that are considered important. An equation can be written that states that the rate of growth of GNP depends upon the rates of growth of the labour force, the capital stock, and other variables. A common procedure is to assume that the influence of the separate inputs is additive—*i.e.*, that the increase in the growth of output caused by increasing the rate of growth of, say, capital is independent of the rate of growth of the labour force. This is the starting point of a great deal of current empirical work that attempts to quantify the importance of different inputs.

Capital and labour

Under certain assumptions, some reasonable and some patently false, it is possible to conclude that what labour and capital receive in the form of wages, profits, and interest is a fair measure of what they contribute to the productive process. Thus in the United States in the period following World War II the share of output going to labour was approximately 79 percent, while the share of output distributed as "profits" was 21 percent. If we assume that these proportions determine how much we should weight the rate of growth of the labour force and of capital respectively in determining their contribution to the rate of growth of output, we must conclude that the relative contribution of capital is slight. Alternatively, we may say that some given percentage increase in the rate of growth of the labour force will have a much larger influence on the rate of growth of output than the same percentage increase in the rate of growth of capital. This is a puzzling result and can be traced to the assumption that the influence of separate inputs is additive.

Quality improvements in the inputs. Much work has been done in an effort to measure the inputs in the productive process more accurately by taking account of improvements in the quality of both labour and capital over time. For example, it has been argued that the amount of a worker's time spent on his formal education is positively related to the income he receives and to his productive contribution. Measuring the number of man-hours worked from one period to the next will not give a true picture of the increase in labour input if the average amount of education received by workers is changing. Man-hour units must be converted to "efficiency" units. Thus if a labour force of 100 workers in the first year all had an eighth-grade education, while 20 years later each member had a 10th-grade education, then measured in efficiency units the labour force had grown. If the length of time spent on formal education increases over time, then the growth of the labour input will be larger if measured in efficiency units. There is, thus, an element of capital in the labour force.

Investment in human capital

Examples of investment in human capital are expenditures on health and on all types of education, including on-the-job training. Expenditures of this sort increase the quality of the labour force and its ability to perform productive tasks. Many economists have argued that technological progress is really nothing but quality improvements in human beings. Some economists take an even broader view and speak of the "production of knowledge" as the clue to technological progress. The production of knowledge is a broad category including outlays on all forms of education, on basic research, and on the more applied type of research associated especially with industry. It is argued that fast-growing industries tend to be those having a high research and development component in their total costs. In addition, firms within an industry that have large research and development budgets tend to experience the most rapid technological progress. The argument is that technical change and improvements must originate in inventions that lead to innovations in the products produced or in the processes whereby existing products are manufactured.

A similar argument applies to the size of the capital stock. It can be maintained that design improvements increase the efficiency of capital goods so that a dollar's worth of machinery purchased today may be much more efficient than a dollar's worth of depreciated machinery purchased yesterday. The rate of growth of the capital stock measured so as to take account of quality improvements will be greater than the rate of growth of the capital stock measured in a way that neglects the differences between "vintages."

Some economists have stressed "economies of scale." For example, if an increase in the use of capital and labour leads to a greater than proportionate increase in output, this is said to result from economies of scale. Economies of scale may arise because an expansion of the market justifies a radical change in productive techniques. These new techniques may be so much more efficient that the returns in the way of increased output are much greater proportionately than the increase in inputs.

Another source of growth and of technical progress in particular has been seen in shifts of demand from low productivity sectors to high productivity sectors, thus causing resources to be reallocated. The most notable movement has been the shift of resources, especially labour, out of agriculture—a traditionally low-productivity sector. Such shifts act to increase the rate of growth of output in ways that cannot be accounted for by simply measuring growth in total inputs. Historically, the allocation of both capital and labour have shifted during the growth process from low productivity sectors to high ones, causing the rate of growth of output to exceed the weighted average of the rates of growth of total inputs.

Entrepreneurship. This historical fact points to an element that has received little attention so far: the influence of entrepreneurship. If the allocation of resources changes during the course of growth and development, it does so under the leadership of an entrepreneurial class. The quality of entrepreneurship is seen by many economists as an important explanation of differences in the rate of technical progress between countries. Decisions must be made somewhere along the line as to whether a new product or process will be introduced. It has been argued that two countries undertaking similar amounts of investment leading to more or less identical rates of growth in the capital stock will not necessarily show the same rate of technical progress. In one country entrepreneurs may be undertaking enterprise investment that has as its aim the introduction of the most advanced types of production techniques, those that will lead to a rapid growth of labour productivity. In the other, because of hesitation or ignorance, the investment program may lead only to marginal changes in productive processes; the resulting growth in labour productivity and GNP will be small. For example, much has been said since World War II about the more aggressive nature of German businessmen as compared to their English counterparts. The emphasis on the role of the entrepreneur in economic growth stems from the theoretical work of the economist Joseph A. Schumpeter, but many others have echoed it.

The play of influences. Much thinking assumes, then, that contributions to output from growth of individual inputs are independent of one another. This assumption allows many growth theorists to conclude that capital investment is relatively unimportant as a growth factor. If there is interaction between the rates of growth of the different inputs, however, then it is possible to draw different conclusions. For example, over time there are likely

Arguments about the significance of capital investment

to be improvements in the quality of capital goods. A machine that requires so much steel and so much labour to manufacture may be twice as productive as an older machine that required the same amount of raw materials and labour in its manufacture. Thus the rate of growth of technical progress and the rate of growth of the capital stock measured in natural units interact. Furthermore, the interaction between technical progress and capital formation is not necessarily in one direction. New knowledge opens up new production possibilities and gives rise to potential increases in technical progress and profits. Or the better educated the labour force, the more adaptable it is likely to be and therefore the better able to cope with new production techniques. At the same time, the higher the rate of growth of capital, the higher will be the growth of incomes and therefore the demand for education. The fact that much of the overall growth of technical progress stems from the transfer of resources and the positive association between the rate of transfer of resources and the rate of growth of the capital stock is another example of interdependence or complementarity between the growth of the inputs. But, again, capital investment undertaken to develop new lines of production will also be dependent on technological progress going on in those areas.

Conventional marginal productivity doctrine argues that as an input such as capital rises relative to labour, the additional output or marginal product that can be attributed to this extra amount of capital will be less than what a unit of capital on the average had been producing before. Marginal productivity doctrine also assumes that each unit of capital is identical with the next. This assumption is the basis for the argument that as more units of capital are utilized in production with a given amount of labour, it will push down the former's marginal product. There is the possibility, however, that additional units of capital may enhance the productivity of existing units: for example, an increase in the amount of capital resources devoted to the development of transportation and distribution may raise the productivity of capital employed, say, in manufacturing. The development of this kind of social overhead capital is certainly a prerequisite for a high return to capital in manufacturing, wholesaling, and retailing.

The analysis can be carried back one more step, to the basic determinants of growth. Economists ask why it is that capital, labour, or technical progress has grown more rapidly in one economy than in another or at one time than at another. Historically, the transition from a subsistence-level, underdeveloped state to a higher-level, developed one has been accompanied by a decline in the death rate followed by a decline in the birth rate. This has the effect of first speeding up the rate of growth of the population and labour force and then reducing it as birth rates fall. Migration can alter this picture, often unpredictably. In the United States, for example, the rate of growth of the population and labour force during the 19th and early 20th centuries was higher than in most other developed countries, mainly because of high rates of immigration. From 1840 to 1930, the native-born U.S. population increased about 600 percent, while the number of those of foreign birth increased 1,300 percent.

The role of government. The differences in rates of growth are often attributed to two factors: government and entrepreneurship. The two are not mutually exclusive. In the early stages of sustained growth, government has often provided the incentives for entrepreneurship to take hold. In some economies the development of transportation, power, and other utilities has been carried out by the government. In others the government has offered financial inducements and subsidies. The land given U.S. railroad developers in the second half of the 19th century is a notable example of the latter. Another important role governments have played in the early stages is to help establish the sort of capital and money markets in which lenders could have confidence. Without financial intermediaries acting as brokers between lenders and business borrowers, it is difficult to envisage economic growth taking place on a sustained and rapid basis.

Government measures in the 19th century

In the 19th century most liberal thinkers held that the main role for government in a developed capitalist system was that of a policeman: to preserve law and order, uphold the sanctity of private property, and give business as much freedom as possible. The Great Depression of the 1930s persuaded many that a laissez-faire system did not automatically provide the necessary incentives to the innovation and risk bearing essential for economic growth. This led to a good deal of writing on the role that governments might play in stimulating growth. Economists have argued that, at the very least, governments can undertake to prevent serious and prolonged recessions. Only in this way can a general business psychology be developed that assumes growth to be the natural course of things, so that investment programs will pay off.

Growth theorists since World War II have gone further, arguing that it is not enough simply to achieve full employment periodically. Some maintain that it is necessary to maintain full employment over an extended period of time if high growth is to result. This argument relates to the earlier point that two economies may experience the same rate of growth of capital but that overall growth and technical progress will proceed at a much more rapid rate in one than in the other because of differences in the quality of new capital goods produced. The term enterprise investment has been used to describe the kind of capital formation that involves innovations and that by building ahead of demand generates rapid rates of growth of productivity or technical progress. But to get such growth, it has been argued, an economy must be run "flat out," at full speed. While this has been subject to some dispute, there is a fairly general consensus that growth will be faster when unemployment fluctuates within a narrow range and at low levels.

Government intervention in expenditure

A variation on this argument is the question of how a government may intervene to determine the distribution of output between those types of expenditure that contribute to growth and those that lead to the immediate satisfaction of consumer demand. Here the choice lies between business investment, research, and education on the one hand and consumption on the other. The larger the first three, the more rapid will be the rate of growth. Governments giving a high priority to growth have various means at their disposal for influencing it. Consumption can and has been constrained through increases in income tax rates. The same is true of other tax rates such as the property tax—the chief revenue source for primary and secondary education in the United States. Tax credit for research and development expenditures is a common method for encouraging business outlays that may lead to innovations. The same method has been used to stimulate business investment outlays. "Easy money" policies on the part of the central bank, whereby the cost of borrowed funds and their availability are indirectly regulated in such a way as to encourage business borrowing, may lead to higher levels of real investment.

The true cost of stimulating growth will always be a temporary cut in current consumption. Only in the future can the economic benefits of the higher investment be realized. By the same token, current consumption can always be enlarged by a neglect of the future. It is even possible for current production to be so biased toward the satisfaction of immediate needs that the productive capacity of an economy slowly declines as capital goods are not replaced. Between the extremes of total neglect of future generations and the paring down of current consumption to a bare subsistence minimum lie an infinite number of possibilities.

THE SOCIAL COST OF GROWTH

The belief that governments should have a large say in choosing the "right" rate of growth has also led some writers to challenge the social and economic value of economic growth in an advanced industrial society. They attribute to growth such undesirable side effects of industrialization as traffic congestion, the increasing pollution of air and water, the despoiling of the landscape, and a general decline in man's ability to enjoy the "real" amenities of life. As has been seen, growth is really a transformation whereby certain industries experience a rise in importance followed by an eventual decline as the market for their output be-

comes relatively saturated. Demand, relatively speaking, moves on to other types of industries and products. All of this naturally implies a reallocation of resources over time. The faster these resources move, other things being equal, the more rapidly can growth and transformation proceed. The argument can be recast in terms of this transformation. A slower rate of growth in per capita consumption will slow down the rate of transfer of resources, but it may also result in a more livable environment. The rate of growth of individual welfare, so measured as to take into account non-consumable amenities, may even be increased. Some argue that in a growth-oriented society wants are created faster than the industrial machine can satisfy them, so that people are more dissatisfied and insecure than they would be if growth were not given such a high value. It is held by some critics that, in modern industrial society, consumption exists for the sake of justifying production rather than production being carried out to satisfy consumer desires. These arguments are a powerful challenge to those who see growth as the most important economic goal of a modern society.

THEORIES OF GROWTH

In discussing theories of growth a distinction must be made between theories designed to explain growth (or the lack of growth) in countries that are already developed and those concerned with countries trapped in circumstances of poverty. Most of what follows will be confined to the former.

The savings–investment equation

As the British economist John Maynard Keynes pointed out in the 1930s, saving and investment are not usually done by the same persons. The desire to save does not necessarily generate investment. If savers attempt to save a larger share of their income than before (thereby consuming less) and if this is not matched by an equal increase in the desire of others to invest, total spending will decline. A natural reaction on the part of business will be to cut back on production, thereby reducing incomes earned in production. The final effect may be a cumulative movement downward as total demand becomes insufficient to employ all of the labour force. This break in the circular flow of income and expenditure suggests the possibility of a capitalist economy alternately experiencing periods of prolonged and severe unemployment (when desired savings at full employment exceed what the economy wishes to invest at full employment) and periods of serious inflation (when the inequality is reversed). This situation had not been the case historically for developed economies until the early 1970s. In the following discussion, some attention will be paid to the ways in which the various theories of growth account for this important historical fact.

Role of the entrepreneur. Modern growth theory can be said to have started with Joseph A. Schumpeter. Unlike most Keynesian or pre-Keynesian theorists, Schumpeter laid primary stress on the role of the entrepreneur, or businessman. It was the quality of his performance that determined whether capital would grow rapidly or slowly and whether this growth would involve innovation and change—*i.e.,* the development of new products and new productive techniques. Differences in growth rates between countries and between different periods in any one country could be traced largely to the quality of entrepreneurship. The latter in turn reflected certain historical and cultural values carried by the business class. Schumpeter also attributed much of the growth of technical progress and of the supply of labour to the entrepreneur. Thus, in more modern terminology, Schumpeter's explanation of why demand and supply have grown more or less at the same rate would be that supply adjusted to demand while demand in turn reflected the activities and investments of the entrepreneur.

Schumpeter believed that capitalism by its very success "sows the seeds of its own destruction." The American economist Alvin H. Hansen argued in the late 1930s that capitalism was in trouble in the United States for other reasons. According to Hansen, the closing of the geographic frontier, the decline in the rate of population growth, and the capital-saving character of recent innovations had all worked to increase the likelihood of stagnation by reducing the need for investment. The savings available in a mature economy would tend to exceed the amount that the economy would want to invest (at levels of full employment) and by progressively larger amounts as time went on. This condition naturally would lead to increasing rates of unemployment as the discrepancy between demand and potential output widened. Hansen's views were very much coloured by the economic conditions of the 1930s. The record of the three decades after World War II did much to overcome the pessimism generated by the Great Depression.

The role of investment. In Keynes's *General Theory,* investment played a key role in that it was presented as the most important factor governing the level of spending in an economy, despite the fact that it typically was only one-fifth to one-sixth of total spending. This paradox can be understood in terms of a concept also developed in the 1930s, the multiplier. The multiplier was the amount by which a change in investment would be multiplied in achieving its final effect on incomes or expenditures. If, for example, investment increases by $10, the extra $10 of expenditures will generate, assuming unemployed resources, an extra $10 of production and subsequently incomes in the form of wages and profit. This increase, however, is hardly the end of the matter since most of the additional incomes earned will be respent on consumer goods. If nine-tenths of any change in income is spent on consumer goods and one-tenth is saved, consumption will increase by $9. But again, one person's expenditures are another person's income, so that incomes now rise by $9 of which $8.1 is respent on consumer goods. The process continues until expenditures, incomes, and production have increased by $100, of which $90 is consumption and $10 the original change in investment. In this case the multiplier is 10.

The negative aspects of investment

But investment may be a source of instability if it is not maintained at a rate sufficient to stimulate demand for the production it is creating. Is there any guarantee that supply or productive capacity will grow at the same rate as demand so that neither excess capacity nor excess demand results? The British economist R.F. Harrod and the American economist E.D. Domar put this question in a very simple mathematical form. In their equations, the rate of growth of supply (*i.e.,* the production function as defined above) is equal to the rate of growth of capital stock. Through investment this capital stock is augmented. The rate of growth of demand depends upon the rate of growth of investment or, more correctly, upon the rate of growth of nonconsumption expenditures. Thus investment affects both demand and supply. But the Harrod–Domar analysis still did not answer the question of what kept the system from becoming increasingly unstable.

Demand and supply. Much contemporary growth theory can be viewed as an attempt to develop a theoretical model that would bring the rate of growth of demand and the rate of growth of supply into line, since a model implying that capitalist systems are inherently unstable would not correspond to the historical facts. Models of growth may be classified according to whether they emphasize adjustments in demand (supply-determined models) or adjustments in supply (demand-determined models). One of the better-known examples of the supply-determined model was developed by the British economist J.R. Hicks. Hicks assumed that the spending propensities of consumers and investors were such as to cause demand to grow at a rate in excess of the rate of growth of maximum output. This assumption meant that during any "boom" the economy would eventually run into a "ceiling" that, while also moving upward, was moving less rapidly than demand. The long-run rate of growth of the economy would be determined by the rate of ascent of the ceiling, which in turn would depend upon supply factors such as the rate of growth of the labour force and the rate of growth of technical progress or productivity. If for some reason these were to grow more rapidly, then output would also grow more rapidly as demand adjusted upward to the more rapid growth of supply.

An example of a demand-determined model of growth is one developed by the American economist J.S. Duesen-

berry. In the Duesenberry model, spending propensities of consumers and investors are such as to generate steady growth in demand. Assume that instead of spending nine-tenths of any change in income on consumer goods, as in the multiplier example above, they choose to spend 0.95. This increase will cause the rate of growth of demand to increase. The question is whether it will also cause the rate of growth of production to increase or whether it will merely result in price increases. If productivity or technical progress responds to a higher rate of growth of demand, as Duesenberry assumes, then production can grow more rapidly. Although in both the Hicks and Duesenberry models demand and supply grow at the same rate, the adjustment mechanisms are entirely different. In the Duesenberry model supply adjusts to demand; in the Hicks model demand adjusts to supply.

Other models of growth also illustrate this distinction between demand-determined and supply-determined growth. The British economist N. Kaldor assumed that there is a mechanism at work generating full employment. Simply stated, in his model an inadequate rate of investment will be offset by shifts in the distribution of income between profits and wages, which will cause consumption to change in a compensating manner so that overall demand is unchanged. While there are important differences between the Hicks and Kaldor models, both can be described as models of supply-determined growth.

Another model of supply-determined growth is that implicit in the traditional neoclassical analysis. The mechanism that adjusts demand to growing supply is the price mechanism, or Adam Smith's "invisible hand" of the market. This type of model assumes a world devoid of monopoly and uncertainty, in which the markets for capital goods and labour are free to adjust quickly so that "markets are always cleared" in the very short run.

Government intervention in demand and supply

A final example of a model of growth that illustrates the problem of adjustment between supply and demand is to be found in the work of the Dutch economist Jan Tinbergen and his followers. In contrast to neoclassical growth models where the market brings about an adjustment of demand to supply, the "target-instrument" models of Tinbergen assume that the government (as in The Netherlands and other European countries) undertakes to regulate demand and supply in an effort to achieve certain targets such as full employment or a predetermined rate of growth. For example, economists are expected to provide the fiscal authorities with a model that approximates the working of the economy and that indicates what will happen if the government, say, does not change its tax and spending programs in the coming period. These forecasts are appraised in terms of what the authorities consider desirable as a matter of social and economic policy. If it appears that unemployment will be too high and the rate of growth too low, the authorities take countermeasures. The government may, for example, cut taxes on corporate profits in order to stimulate investment. If investment is excessive and there is danger of inflation, the government may take other measures to reduce aggregate demand such as cutting its expenditures. This type of planning procedure has been tried with varying degrees of success. Sweden and The Netherlands are prominent examples of attempts to offset fluctuations in private spending so as to realize full employment and growth. It should be noted that these models do not fit neatly into the demand-determined or supply-determined classification. In the example just given, both the rate of growth of demand and the rate of growth of supply are effectively determined by the fiscal authorities.

Economic stagnation. The rise in unemployment rates and the slowdown in growth rates of GNP and per capita incomes throughout the capitalist world beginning in the early 1970s is clearly a case where demand and supply did not grow at similar rates. Many economists turned their attention to developing theories to explain this prolonged period of stagnation. A common theme in much of their work was the adverse effects of high unemployment and low utilization of the capital stock on investment and, therefore, on productivity growth.

The high unemployment rates for labour and capital are initially traced to policies restricting aggregate demand that were pursued by monetary and fiscal authorities from the first half of the 1970s. This policy response was widely interpreted by economists as an effort by the authorities to reduce inflation rates that had begun to accelerate in the latter 1960s. The continued use of restrictive policies is then related to fear on the part of the authorities that any attempt to restimulate their economies would merely bring back inflation.

Tighter labour markets resulting from any such stimulative policies are seen to increase the bargaining power of labour, thereby leading to larger wage demands and settlements that in turn feed into prices, causing price inflation to accelerate. This leads to yet higher wage demands in order to protect real wages and thus an explosive wage–price spiral. In addition, more stimulative aggregate demand policies are perceived to result in balance of payments difficulties at existing exchange rates. But any attempt to avoid larger payments deficits by reducing the exchange rate leads to the "importation" of inflation through higher prices of imported goods. The result of such considerations is reluctance of the authorities to attempt to create full employment through stimulative policies.

Inflation and growth

What emerges from these theories is a chain of causation that describes the way in which, in the period since World War II, inflation and growth have become causally connected through the responses of governments to actual and anticipated inflationary pressures. Inflation and the fear of inflation lead to slow growth and high unemployment because the inability of governments to bring inflation under control at full employment by other means—*e.g.*, an income policy—constrains governments to implement restrictive policies to combat or forestall inflationary pressures. Such responses lead, as they did in the early 1970s, not only to high rates of unemployment of capital and labour but also to low rates of investment and productivity growth. Stagnation is the result, and such a scenario is a likely prospect for capitalism in the future.

Foreign trade. Little has been said about foreign trade. Yet growth in most economies is very much dependent upon imports and the ability to export in order to pay for imports. The fact that some economies recovered relatively quickly from World War II and grew much more rapidly in the postwar period than others has stimulated a great deal of comparative analysis in growth theory. The exceptionally high growth rates in Japan and Germany compared to the general sluggishness of the British economy are related to foreign trade. Economists have pointed to the periodic balance of payments crises experienced by Britain and the lack of such crises in Germany. During a boom, as incomes rise the demand for imports will rise also as a natural feature of prosperity. But if exports do not also rise at the same time, the authorities may be forced to take fiscal or monetary countermeasures and slow down the economy in an effort to bring imports and exports back into balance. Or exports may fail to grow sufficiently because labour costs are rising very rapidly and pushing up prices of exports faster than in competing countries.

A policy of encouraging growth has the effect of keeping the demand for imports high and making labour markets tight, thereby tending to push up money wage rates. At the same time, such a policy also tends to encourage innovations and investment projects that are very productive, particularly if the demand pressures are sustained. A "stop" policy naturally has just the opposite effects, both good and bad from the point of view of a country's balance of payments. The question is which policy will in the long run result in less rapidly rising costs and prices. Many writers have argued that if demand pressures are maintained the response or adjustment of productivity and therefore of supply to these pressures will be such that the country will soon find itself in a more competitive position. Running an economy "flat out," however, is likely to cause a short-run balance of payments crisis and lead to devaluation of currency.

Mathematical growth theories. In addition to the theories discussed above, a large body of literature has developed involving abstract mathematical models. Because this field of analysis is so technical, only a general picture

of the kinds of problems and questions discussed can be given. First, a set of equations is drawn up describing what the model builder feels are the important relations between economic variables such as output, capital, investment, and consumption. These equations must relate economic variables to one another at different points in time: for example, output last year determines consumption this year, which in turn helps to determine output this year and therefore consumption and output next year. It is possible to work out the movements of the variables over as long a period as desired. At the centre of much of this analysis is the concept of a steady-state rate of growth: one in which all the economic variables contained in the set of equations grow at the same constant rate equal to the growth of the labour force.

Optimal growth models

A related class of studies attempts to take account of the welfare of workers and consumers in the maximization of growth. These "optimal growth" models seek to maximize consumer satisfaction over time. In a model such as this the solution will not be the highest possible growth rate but one that will maximize the welfare of consumers. The importance of such models for planners would seem to depend on the realism of their assumptions as to consumer desires and technology.

Model building and theorizing about growth has proceeded on various levels of abstraction. Some of the work is of little practical value, in the sense that its explanatory value is negligible. Such studies, however, may stimulate other work that is helpful in an understanding of the growth process. Some models, while realistic, are not applicable to all economies. Thus, a model that neglects international trade is of little use to a European economist trying to understand the more basic causes of differences in growth rates between countries. (J.L.C.)

Economic development

Economic development first became a major concern after World War II. As the era of European colonialism ended, many former colonies and other countries with low living standards came to be termed underdeveloped countries, to contrast their economies with those of the developed countries, which were understood to be Canada, the United States, those of western Europe, most eastern European countries, the then Soviet Union, Japan, South Africa, Australia, and New Zealand. As living standards in most poor countries began to rise in subsequent decades, they were renamed the developing countries.

There is no universally accepted definition of what a developing country is; neither is there one of what constitutes the process of economic development. Developing countries are usually categorized by a per capita income criterion, and economic development is usually thought to occur as per capita incomes rise. A country's per capita income (which is almost synonymous with per capita output) is the best available measure of the value of the goods and services available, per person, to the society per year. Although there are a number of problems of measurement of both the level of per capita income and its rate of growth, these two indicators are the best available to provide estimates of the level of economic well-being within a country and of its economic growth.

It is well to consider some of the statistical and conceptual difficulties of using the conventional criterion of underdevelopment before analyzing the causes of underdevelopment. The statistical difficulties are well known. To begin with, there are the awkward borderline cases. Even if analysis is confined to the underdeveloped and developing countries in Asia, Africa, and Latin America, there are rich oil countries that have per capita incomes well above the rest but that are otherwise underdeveloped in their general economic characteristics. Second, there are a number of technical difficulties that make the per capita incomes of many underdeveloped countries (expressed in terms of an international currency, such as the U.S. dollar) a very crude measure of their per capita real income. These difficulties include the defectiveness of the basic national income and population statistics, the inappropriateness of the official exchange rates at which the national incomes in terms of the respective domestic currencies are converted into the common denominator of the U.S. dollar, and the problems of estimating the value of the noncash components of real incomes in the underdeveloped countries. Finally, there are conceptual problems in interpreting the meaning of the international differences in the per capita income levels.

Difficulties of measurement

Per capita income comparisons

Although the difficulties with income measures are well established, measures of per capita income correlate reasonably well with other measures of economic well-being, such as life expectancy, infant mortality rates, and literacy rates. Other indicators, such as nutritional status and the per capita availability of hospital beds, physicians, and teachers, are also closely related to per capita income levels. While a difference of, say, 10 percent in per capita incomes between two countries would not be regarded as necessarily indicative of a difference in living standards between them, actual observed differences are of a much larger magnitude. India's per capita income, for example, was estimated at $270 in 1985. In contrast, Brazil's was estimated to be $1,640, and Italy's was $6,520. While economists have cited a number of reasons why the implication that Italy's living standard was 24 times greater than India's might be biased upward, no one would doubt that the Italian living standard was significantly higher than that of Brazil, which in turn was higher than India's by a wide margin.

The interpretation of a low per capita income level as an index of poverty in a material sense may be accepted with two qualifications. First, the level of material living depends not on per capita income as such but on per capita consumption. The two may differ considerably when a large proportion of the national income is diverted from consumption to other purposes; for example, through a policy of forced saving. Second, the poverty of a country is more faithfully reflected by the representative standard of living of the great mass of its people. This may be well below the simple arithmetic average of per capita income or consumption when national income is very unequally distributed and there is a wide gap in the standard of living between the rich and the poor.

The usual definition of a developing country is that adopted by the World Bank: "low-income developing countries" in 1985 were defined as those with per capita incomes below $400; "middle-income developing countries" were defined as those with per capita incomes between $400 and $4,000. To be sure, countries with the same per capita income may not otherwise resemble one another: some countries may derive much of their incomes from capital-intensive enterprises, such as the extraction of oil, whereas other countries with similar per capita incomes may have more numerous and more productive uses of their labour force to compensate for the absence of wealth in resources. Kuwait, for example, was estimated to have a per capita income of $14,480 in 1985, but 50 percent of that income originated from oil. In most regards, Kuwait's economic and social indicators fell well below what other countries with similar per capita incomes had achieved. Centrally planned economies are also generally regarded as a separate class, although China and North Korea are universally considered developing countries. A major difficulty is that prices serve less as indicators of relative scarcity in centrally planned economies and hence are less reliable as indicators of the per capita availability of goods and services than in market-oriented economies.

Estimates of percentage increases in real per capita income are subject to a somewhat smaller margin of error than are estimates of income levels. While year-to-year changes in per capita income are heavily influenced by such factors as weather (which affects agricultural output, a large component of income in most developing countries), a country's terms of trade, and other factors, growth rates of per capita income over periods of a decade or more are strongly indicative of the rate at which average economic well-being has increased in a country.

ECONOMIC DEVELOPMENT AS AN OBJECTIVE OF POLICY

Motives for development. The field of development economics is concerned with the causes of underdevelopment

and with policies that may accelerate the rate of growth of per capita income. While these two concerns are related to each other, it is possible to devise policies that are likely to accelerate growth (through, for example, an analysis of the experiences of other developing countries) without fully understanding the causes of underdevelopment.

Studies of both the causes of underdevelopment and of policies and actions that may accelerate development are undertaken for a variety of reasons. There are those who are concerned with the developing countries on humanitarian grounds; that is, with the problem of helping the people of these countries to attain certain minimum material standards of living in terms of such factors as food, clothing, shelter, and nutrition. For them, low per capita income is the measure of the problem of poverty in a material sense. The aim of economic development is to improve the material standards of living by raising the absolute level of per capita incomes. Raising per capita incomes is also a stated objective of policy of the governments of all developing countries. For policymakers and economists attempting to achieve their governments' objectives, therefore, an understanding of economic development, especially in its policy dimensions, is important. Finally, there are those who are concerned with economic development either because they believe it is what people in developing countries want or because they believe that political stability can be assured only with satisfactory rates of economic growth. These motives are not mutually exclusive. Since World War II many industrial countries have extended foreign aid to developing countries for a combination of humanitarian and political reasons.

Growth for stability

Those who are concerned with political stability tend to see the low per capita incomes of the developing countries in relative terms; that is, in relation to the high per capita incomes of the developed countries. For them, even if a developing country is able to improve its material standards of living through a rise in the level of its per capita income, it may still be faced with the more intractable subjective problem of the discontent created by the widening gap in the relative levels between itself and the richer countries. (This effect arises simply from the operation of the arithmetic of growth on the large initial gap between the income levels of the developed and the underdeveloped countries. As an example, an underdeveloped country with a per capita income of $100 and a developed country with a per capita income of $1,000 may be considered. The initial gap in their incomes is $900. Let the incomes in both countries grow at 5 percent. After one year, the income of the underdeveloped country is $105, and the income of the developed country is $1,050. The gap has widened to $945. The income of the underdeveloped country would have to grow by 50 percent to maintain the same absolute gap of $900.) Although there was once in development economics a debate as to whether raising living standards or reducing the relative gap in living standards was the true desideratum of policy, experience during the 1960–80 period convinced most observers that developing countries could, with appropriate policies, achieve sufficiently high rates of growth both to raise their living standards fairly rapidly and to begin closing the gap. (H.My./A.O.K.)

Compound growth of income gap

The impact of discontent. Although concern over the question of a subjective sense of discontent among the underdeveloped and developing countries has waxed and waned, it has never wholly disappeared. The underdeveloped countries' sense of dissatisfaction and grievance arises not only from measurable differences in national incomes but also from the less easily measurable factors, such as their reaction against the colonial past and their complex drives to raise their national prestige and achieve equality in the broadest sense with the developed countries. Thus, it is not uncommon to find their governments using a considerable proportion of their resources in prestige projects, ranging from steel mills, hydroelectric dams, universities, and defense expenditure to international athletics. These symbols of modernization may contribute a nationally shared satisfaction and pride but may or may not contribute to an increase in the measurable national income. Second, it is possible to argue that in many cases the internal gap in incomes within individual underdeveloped countries may be a more potent source of the subjective level of discontent than the international gap in income. Faster economic growth may help to reduce the internal economic disparities in a less painful way, but it must be remembered that faster economic growth also tends to introduce greater disruption and the need for making bigger readjustments in previous ways of life and may thus increase the subjective sense of frustration and discontent. Finally, it is difficult to establish that the subjective problem of discontent will bear a simple and direct relationship to the size of the international gap in incomes. Some of the apparently most discontented countries are to be found in Latin America, where the per capita incomes are generally higher than in Asia and Africa. A skeptic can turn the whole approach to a reductio ad absurdum by pointing out that even the developed countries with their high and rising levels of per capita income have not been able to solve the subjective problem of discontent and frustration among various sections of their population.

Two conclusions may be drawn from the above points. First, the subjective problem of discontent in the underdeveloped countries is a genuine and important problem in international relations. But economic policy acting on measurable economic magnitudes can play only a small part in the solution of what essentially is a problem in international politics. Second, for the narrower purpose of economic policy there is no choice but to fall back on the interpretation of the low per capita incomes of the underdeveloped countries as an index of their poverty in a material sense. This can be defended by explicitly adopting the humanitarian value judgment that the underdeveloped countries ought to give priority to improving the material standards of living of the mass of their people. But, even if this value judgment is not accepted, the conventional measure of economic development in terms of a rise in per capita income still retains its usefulness. The governments of the underdeveloped countries may wish to pursue other, nonmaterial goals, but they could make clearer decisions if they knew the economic cost of their decisions. The most significant measure of this economic cost can be expressed in terms of the foregone opportunity to raise the level of per capita income. (H.My.)

Toward a workable criterion

A SURVEY OF DEVELOPMENT THEORIES

The hypothesis of underdevelopment. If the underdeveloped countries are merely low-income countries, why call them underdeveloped? The use of the term underdeveloped in fact rests on a general hypothesis on which the whole subject matter of development economics is based. According to this hypothesis, the existing differences in the per capita income levels between the developed and the underdeveloped countries cannot be accounted for purely in terms of differences in natural conditions beyond the control of man and society. That is to say, the underdeveloped countries are underdeveloped because, in some way or another, they have not yet succeeded in making full use of their potential for economic growth. This potential may arise from the underdevelopment of their natural resources, or their human resources, or from the "technological gap." More generally, it may arise from the underdevelopment of economic organization and institutions, including the network of the market system and the administrative machinery of the government. The general presumption is that the development of this organizational framework would enable an underdeveloped country to make a fuller use not only of its domestic resources but also of its external economic opportunities, in the form of international trade, foreign investment, and technological and organizational innovations.

Development thought after World War II. After World War II a number of developing countries attained independence from their former colonial rulers. One of the common claims made by leaders of independence movements was that colonialism had been responsible for perpetuating low living standards in the colonies. Thus economic development after independence became an objective of policy not only because of the humanitarian desire to raise living standards but also because political promises had been made, and failure to make progress

Growth as a political desideratum

toward development would, it was feared, be interpreted as a failure of the independence movement. Developing countries in Latin America and elsewhere that had not been, or recently been, colonies took up the analogous belief that economic domination by the industrial countries had thwarted their development, and they, too, joined the quest for rapid growth.

At that early period, theorizing about development, and about policies to attain development, accepted the assumption that the policies of the industrial countries were to blame for the poverty of the developing countries. Memories of the Great Depression, when developing countries' terms of trade had deteriorated markedly, producing sharp reductions in per capita incomes, haunted many policymakers. Finally, even in the developed countries, the Keynesian legacy attached great importance to investment.

In this milieu, it was thought that a "shortage of capital" was the cause of underdevelopment. It followed that policy should aim at an accelerated rate of investment. Since most countries with low per capita incomes were also heavily agricultural (and imported most of the manufactured goods consumed domestically), it was thought that accelerated investment in industrialization and the development of manufacturing industries to supplant imports through "import substitution" was the path to development. Moreover, there was a fundamental distrust of markets, and a major role was therefore assigned to government in allocating investments. Distrust of markets extended especially to the international economy.

Experience with development changed perceptions of the process and of the policies affecting it in important ways. Nonetheless, there are significant elements of truth in some of the earlier ideas, and it is important to understand the thinking underlying them. (H.My./A.O.K.)

Growth economics and development economics. Development economics may be contrasted with another branch of study, called growth economics, which is concerned with the study of the long-run, or steady-state, equilibrium growth paths of the economically developed countries, which have long overcome the problem of initiating development.

Growth theory assumes the existence of a fully developed modern capitalist economy with a sufficient supply of entrepreneurs responding to a well-articulated system of economic incentives to drive the growth mechanism. Typically, it concentrates on macroeconomic relations, particularly the ratio of savings to total output and the aggregate capital–output ratio (that is, the number of units of additional capital required to produce an additional unit of output). Mathematically, this can be expressed (the Harrod–Domar growth equation) as follows: the growth in total output (g) will be equal to the savings ratio (s) divided by the capital–output ratio (k); *i.e.*, $g = \frac{s}{k}$. Thus, suppose that 12 percent of total output is saved annually and that three units of capital are required to produce an additional unit of output: then the rate of growth in output is $\frac{12}{3}\% = 4\%$ per annum. This result is obtained from the basic assumption that whatever is saved will be automatically invested and converted into an increase in output on the basis of a given capital–output ratio. Since a given proportion of this increase in output will be saved and invested on the same basis, a continuous process of growth is maintained.

Misapplication of growth theory

Growth theory, particularly the Harrod–Domar growth equation, has been frequently applied or misapplied to the economic planning of a developing country. The planner starts from a desired target rate of growth of perhaps 4 percent. Assuming a fixed overall capital–output ratio of, say, 3, it is then asserted that the developing country will be able to achieve this target rate of growth if it can increase its savings to 3×4 percent $= 12$ percent of its total output. The weakness of this type of exercise arises from the assumption of a fixed overall capital–output ratio, which assumes away all the vital problems affecting the developing country's capacity to absorb capital and invest its saving in a productive manner. These problems include the central problem of the efficient allocation of available savings among alternative investment opportunities and the associated organizational and institutional problems of encouraging the growth of a sufficient supply of entrepreneurs; the provision of appropriate economic incentives through a market system that correctly reflects the relative scarcities of products and factors of production; and the building up of an organizational framework that can effectively implement investment decisions in both the private and the public sectors. Such problems, which generally affect the developing country's absorptive capacity for capital and a number of other inputs, constitute the core of development economics. Development economics is needed precisely because the assumptions of growth economics, based as they are on the existence of a fully developed and well-functioning modern capitalist economy, do not apply.

The developing and underdeveloped countries are a very mixed collection of countries. They differ widely in area, population density, and natural resources. They are also at different stages in the development of market and financial institutions and of an effective administrative framework. These differences are sufficient to warn against wide-sweeping generalizations about the causes of underdevelopment and all-embracing theoretical models of economic development. But when development economics first came into prominence in the 1950s, there were powerful intellectual and political forces propelling the subject toward such general theoretical models of development and underdevelopment. First, many writers who popularized the subject were frankly motivated by a desire to persuade the developed countries to give more economic aid to the underdeveloped countries, on grounds ranging from humanitarian considerations to considerations of cold-war strategy. Second, there was the reaction of the newly independent underdeveloped countries against their past "colonial economic pattern," which they identified with free trade and primary production for the export market. These countries were eager to accept general theories of economic development that provided a rationalization for their deep-seated desire for rapid industrialization. Third, there was a parallel reaction, at the academic level, against older economic theory, with its emphasis on the efficient allocation of scarce resources and a striving after new and "dynamic" approaches to economic development.

Reasons for the change in theoretical models

All of these forces combined to produce a crop of theoretical approaches that soon developed into a fairly fixed orthodoxy with its characteristic emphasis on "crash" programs of investment in both material and human capital, on domestic industrialization, and on government economic planning as the standard ingredients of development policy. These new theories have continued to have a considerable influence on the conventional wisdom in development economics, although in retrospect most of them have turned out to be partial theories. A broad survey of these theories, under three main heads, is given below. It is particularly relevant to the debate over whether the underdeveloped countries should seek economic development through domestic industrialization or through international trade. The limitations of these new theories—and how they led to a gradual revival of a more pragmatic approach to development problems, which falls back increasingly on the older economic theory of efficient allocation of resources—are subsequently traced.

The missing-component approach. First, there are the theories that regard the shortage of some strategic input (such as the supply of savings, foreign exchange, or technical skills) as the main cause of underdevelopment. Once this missing component was supplied—say, by external economic aid—it was believed that economic development would follow in a predictable manner based on fixed quantitative relationships between input and output. The overall capital–output ratio, mentioned above, is the most well-known of these fixed technical coefficients. But similar fixed coefficients have been assumed between the foreign-exchange requirements and total output and between the input of skilled manpower and output.

(H.My.)

Shortage of savings. Given the broad relationship between capital accumulation and economic growth es-

tablished in growth theory, it was plausible for growth theorists and development economists to argue that the developing countries were held back mainly by a shortage in the supply of capital. These countries were then saving only 5–7 percent of their total product, and it was manifest (and it remains true) that satisfactory growth cannot be supported by so low a level of investment. It was therefore thought that raising the savings ratio to 10–12 percent was the central problem for developing countries. Early development policy therefore focused on raising resources for investment. Steps toward this end were highly successful in most developing countries, and savings ratios rose to the 15–25 percent range. However, growth rates failed even to approximate the savings rates, and theorists were forced to search for other explanations of differences in growth rates.

Aid and growth

It has become increasingly clear that there can be much wastage of capital resources in the developing countries for various reasons, such as wrong choice of investment projects, inefficient implementation and management of these projects, and inappropriate pricing and costing of output. These faults are particularly noticeable in public-sector investment projects and are one of the reasons why the Pearson Commission Report of the International Bank for Reconstruction and Development (1969) found that "the correlation between the amounts of aid received in the past decades and the growth performance is very weak." But even in the private sector there may be a considerable distortion in the direction of investment induced by policies designed to encourage development. Thus, in most underdeveloped countries, a considerable part of private expansion investment, both foreign and domestic, has been diverted into the expansion of the manufacturing sector, catering to the domestic market through various inducements, including tariff protection, tax holidays, cheap loans, and generous foreign-exchange allocations granting the opportunity to import capital goods cheaply at overvalued exchange rates. As a consequence, there developed a very considerable amount of excess capacity in the manufacturing sector of the underdeveloped countries pursuing such policies.

Foreign-exchange shortage. In the 1950s most developing countries were primary commodity exporters, relying on crops and minerals for the bulk of their foreign-exchange earnings through exports, and importing a large number of manufactured goods. The experience of colonialism, and the distrust of the international economy that it engendered, led policymakers in most developing countries to adopt a policy of import substitution. This policy was intended to promote industrialization by protecting domestic producers from the competition of imports. Protection, in the form of high tariffs or the restriction of imports through quotas, was applied indiscriminately, often to inherently high-cost industries that had no hope of ever becoming internationally competitive. Also, after the early stages of import substitution, protected new industries tended to be very intensive in the use of capital and especially of imported capital goods.

The import-substitution approach defined "industrialization" rather narrowly as the expansion of the modern manufacturing sector based on capital-intensive technology. Capital was therefore identified with durable capital equipment in the form of complex machinery and other inputs that the underdeveloped countries were not able to produce domestically. Thus, foreign-exchange requirements were calculated on the basis of the fixed technical input-output coefficients of the manufacturing sector.

With high levels of protection for domestic industry, and with exchange rates that were often maintained at unrealistic levels (usually in an effort to make imported capital goods "cheap"), the experience of most developing countries was that export earnings grew relatively slowly. The simultaneously sharp increase in demand for imported capital goods (and for raw materials and replacement parts as well) resulted in unexpectedly large increases in imports. Most developing countries found themselves with critical foreign-exchange shortages and were forced to reduce imports in order to cut their current-account deficits to manageable proportions.

The cutbacks in imports usually resulted in reduced growth rates, if not recessions. This result led to the view that economic stagnation was caused primarily by a shortage of foreign exchange with which to buy essential industrial inputs. But over the longer term the growth rates of countries that continued to protect their domestic industries heavily not only stagnated but declined sharply. Contrasting the experience of countries that persisted in policies of import substitution with those that followed alternative policies (see below) subsequently demonstrated that foreign-exchange shortage was a barrier to growth only within the context of the protectionist policies adopted and was not inherently a barrier to the development process itself. (H.My./A.O.K.)

Education and human capital in development. As it became apparent that the physical accumulation of capital was not by itself the key to development, many analysts turned to a lack of education and skills among the population as being a crucial factor in underdevelopment. If education and skill are defined as everything that is required to raise the productivity of the people in the developing countries by improving their skills, enterprise, initiative, adaptability, and attitudes, this proposition is true but is an empty tautology. However, the need for skills and training was first formulated in terms of specific skills and educational qualifications that could be supplied by crash programs in formal education. The usual method of manpower planning thus started from a target rate of expansion in output and tried to estimate the numbers of various types of skilled personnel that would be required to sustain this target rate of economic growth on the basis of an assumed fixed relationship between inputs of skill and national output.

This approach was plausible enough in many developing countries immediately after their political independence, when there were obvious gaps in various branches of the administrative and technical services. But most countries passed through this phase rather quickly. In the meantime, as the result of programs in education expansion, their schools and colleges began producing large numbers of fresh graduates at much faster rates than their general rate of economic growth could supply suitable new jobs for. This created a growing problem of educated unemployment. An important factor behind the rapid educational expansion was the expectation that after graduation students would be able to obtain well-paying white-collar jobs at salary levels many times the prevailing per capita income of their countries. Thus, the underdeveloped countries' inability to create jobs to absorb their growing armies of graduates created an explosive element in what came to be called the revolution of expectations.

Educated unemployment

It is possible to see a close parallelism between the narrow concept of industrialization as the expansion of the manufacturing sector and the narrow concept of education as the academic and technical qualifications that can be supplied by the expansion of the formal educational system. If a broader concept of education, relevant for economic development, is needed, it is necessary to seek it in the pervasive educational influence of the economic environment as a whole on the learning process of the people of the underdeveloped countries. This is a complex process that depends on, among other less easily analyzable things, the system of economic incentives and signals that can mold the economic behaviour of the people of the underdeveloped countries and affect their ability to make rational economic decisions and their willingness to introduce or adapt to economic changes. Unfortunately, the economic environment in many underdeveloped countries is dominated by a network of government controls that tend not to be conducive to such ends.

Surplus resources and disguised unemployment. Two theories emphasized the existence of surplus resources in developing countries as the central challenge for economic policy. The first concentrated on the countries with relatively abundant natural resources and low population densities and argued that a considerable amount of both surplus land and surplus labour might still exist in these countries because of inadequate marketing facilities and lack of transport and communications. Economic devel-

opment was pictured as a process whereby these underutilized resources of the subsistence sector would be drawn into cash production for the export market. International trade was regarded as the chief market outlet, or vent, for the surplus resources. The second theory was concerned with the thickly populated countries and the possibility of using their surplus labour as the chief means of promoting economic development. According to this theory, because of heavy population pressure on land, the marginal product of labour (that is, the extra output derived from the employment of an extra unit of labour) was reduced to zero or to a very low level. But the people in the subsistence sector were able to enjoy a certain customary minimum level of real income because the extended-family system of the rural society shared the total output of the family farm among its members. A considerable proportion of labour in the traditional agricultural sector was thus thought to contribute little or nothing to total output and to really be in a state of disguised unemployment. By this theory, the labour might be drawn into other uses without any cost to society.

The concept of disguised unemployment

It is necessary to clear up a number of preliminary points about the concept of disguised unemployment before considering its applications. First, it is highly questionable whether the marginal product of labour is actually zero even in densely populated countries such as India or Pakistan. Even in these countries, with existing agricultural methods, all available labour is needed in the peak seasons, such as harvest. The most important part of disguised unemployment is thus what may be better described as seasonal unemployment during the off-seasons. The magnitude of this seasonal unemployment, however, depends not so much on the population density on land as on the number of crops cultivated on the same piece of land through the year. There is thus little seasonal unemployment in countries such as Taiwan or South Korea, which have much higher population densities than India, because improved irrigation facilities enable them to grow a succession of crops on the same land throughout the year. But there may be considerable seasonal unemployment even in sparsely populated countries growing only one crop a year.

Seasonal unemployment

The main weakness in the proposal to use disguised unemployment for the construction of major social-overhead-capital projects arises from an inadequate consideration of the problem of providing the necessary subsistence fund to maintain the workers during what may be a considerably long waiting period before these projects yield consumable output. This may be managed somehow for small-scale local-community projects when the workers are maintained in situ by their relatives. But when it is proposed to move a large number of surplus workers away from their home villages for major construction projects taking a considerable time to complete, the problem of raising a sufficient subsistence fund to maintain the labour becomes formidable. The only practicable way of raising such a subsistence fund is to encourage voluntary saving and the expansion of a marketable surplus of food that can be purchased with the savings to maintain the workers. The mere existence of disguised unemployment does not in any way ease this problem. (H.My.)

Role of governments and markets. In earlier thinking about development, it was assumed that the market mechanisms of developed economies were so unreliable in developing economies that governments had to assume central responsibility for economic activity. This was to be done through economic planning for the entire economy (see below *Planning in developing countries*), which in turn would be implemented by active government participation in the economy and pervasive controls over all private-sector economic activity. Government participation took many forms: Public-sector enterprises were established to manufacture many commodities, including steel, machine tools, fertilizers, heavy chemicals, and even textiles and clothing; government marketing boards assumed monopoly power over the purchase and sale of many agricultural commodities; and government agencies became the sole importers of a variety of goods, and they often became exporters as well. Controls over private-sector activity were even more extensive: Price controls were established for many commodities; import licensing procedures eliminated the importing of commodities not given priority in official plans; investment licenses were required before factories could be expanded; capacity licenses regulated maximum permissible outputs; and comprehensive regulations governed the conditions of employment of workers.

Distrust of market mechanisms

The consequence, frequently, was that indigenous entrepreneurs often found it more financially rewarding to devote their energies and ingenuity to the task of procuring the necessary government import licenses and other permits and exploiting the loopholes in government regulations than to the problem of raising the efficiency and productivity of resources. For public-sector enterprises, political pressures often resulted in the employment of many more persons than could be productively used and in other practices conducive to extremely high-cost and inefficient operations. The consequent fiscal burden diverted resources that might otherwise have been used for investment, while the inefficient use of resources dampened growth rates.

Related to the belief in market failure and in the necessity for government intervention was the view that the efficiency of the price mechanism in developing countries was very small. This was reflected in the view of foreign-exchange shortage, already discussed, in which it was thought that there are fixed relationships between imported capital and domestic expansion. It was also reflected in the view that farmers are relatively insensitive to relative prices and in the belief that there are few entrepreneurs in developing countries.

LESSONS FROM DEVELOPMENT EXPERIENCE

By the end of the 1950s the experience gained from efforts to promote economic development showed great differences among developing countries. Some had broken away relatively quickly from the import-substitution, government-control and -ownership pattern that had been the early development wisdom. Others persisted with the same policies for several decades. A great deal was learned from the experiences of different developing countries.

The importance of agriculture. Despite early emphasis on industrialization through import substitution, a first major lesson of postwar experience was that there is a close connection between the rate of growth in the output of the agricultural sector and the general rate of economic development. The high rates of economic growth are associated with rapid expansion of agricultural output and low rates of economic growth with the slow growth of agriculture. This is (in hindsight, at least) to be expected, since agriculture forms a large part of the total domestic product and of the exports of the developing countries. What is more interesting is that the expansion of agricultural output was by no means confined to those countries with an abundant supply of unused land to be brought under cultivation. Taiwan and South Korea, with some of the highest population densities in the world, were able to expand their agricultural output rapidly by a vigorous pursuit of appropriate policies. These included the provision of adequate irrigation facilities, enabling a succession of crops to be grown on the same piece of land throughout the year; the use of high-yielding seeds and fertilizers, which raised the yields per acre in a dramatic fashion; provision of adequate incentives for producers by setting producer prices at reasonable levels; and improvements in credit and marketing facilities and a general improvement in the economic organization of the agricultural sector. Agricultural development is important because it raises the incomes of the mass of the people in the countryside; in addition, it increases the size of the domestic market for the manufacturing sector and reduces internal economic disparities between the urban centres and the rural districts.

The role of exports. A second conclusion to be drawn from experience is the close connection between export expansion and economic development. The high-growth countries were characterized by rapid expansion in exports. Here again it is important to note that export ex-

pansion was not confined to those countries fortunate in their natural resources, such as the oil-exporting countries. Some of the developing countries were able to expand their exports in spite of limitations in natural resources by initiating economic policies that shifted resources from inefficient domestic manufacturing industries to export production. Nor was export expansion from the developing countries confined to primary products. There was very rapid expansion of exports of labour-intensive manufactured goods. This phenomenon occurred not only in the extremely rapidly growing, newly industrialized countries (NICs)—Singapore, Hong Kong, South Korea, and Taiwan—but also from other developing countries including Brazil, Argentina, and Turkey. Countries that adopted export-oriented development strategies (of which the most notable were the NICs) experienced extremely high rates of growth that were regarded as unattainable in the 1950s and 1960s. They were also able to maintain their growth momentum during periods of worldwide recession better than were the countries that maintained their import substitution policies.

The newly industrialized countries (NICs)

Analysts have pointed to a number of reasons why the export-oriented growth strategy seems to deliver more rapid economic development than the import substitution strategy. First, a developing country able to specialize in producing labour-intensive commodities uses its comparative advantage in the international market and is also better able to use its most abundant resource—unskilled labour. The experience of export-oriented countries has been that there is little or no disguised unemployment once labour-market regulations are dismantled and incentives are created for individual firms to sell in the export market. Second, most developing countries have such small domestic markets that efforts to grow by starting industries that rely on domestic demand result in uneconomically small, inefficient enterprises. Moreover, those enterprises will typically be protected from international competition and the incentives it provides for efficient production techniques. Third, an export-oriented strategy is inconsistent with the impulse to impose detailed economic controls; the absence of such controls, and their replacement by incentives, provides a great stimulus to increases in output and to the efficiency with which resources are employed. The increasing capacity of a developing country's entrepreneurs to adapt their resources and internal economic organization to the pressures of world-market demand and international competition is a very important connecting link between export expansion and economic development. It is important in this connection to stress the educative effect of freer international trade in creating an environment conducive to the acceptance of new ideas, new wants, and new techniques of production and methods of organization from abroad.

The negative effect of controls. Another major lesson that was learned is that poor people are, if anything, more responsive to incentives than rich people. Nominal exchange rates that are pegged without regard to domestic inflation have strong negative effects on incentives to export; producer prices for agricultural goods that are set as a small fraction of their world market price constitute a significant disincentive to agricultural production; and controls on prices and investment serve as significant deterrents to economic activity. Indeed, in most environments, controls lead to "rent-seeking" behaviour, in which resources are diverted from productive activity and instead are used to try to win import licenses, or to get the necessary bureaucratic permissions. In addition, in many countries, "parallel," or black, markets emerged, which diverted resources from activities in the official sector. In some countries, legal exports diminished sharply as smuggling and underinvoicing intensified in response to increasing discrepancies between the official exchange rate and the black-market rate.

Incentives and disincentives

The importance of appropriate incentives. As a corollary to the lesson that controls may strongly divert economic activity from an efficient allocation of resources, it became increasingly evident that inappropriate incentives can adversely affect economic behaviour. The response of agricultural supply to increases in producer prices is considerably stronger than was earlier believed. Likewise, individuals respond to incentives with respect to their education and training. Thus, much of the overinvestment in education referred to earlier came to be seen as the result of artificially inflated wages for university graduates in the public sector and of the fact that university education was virtually free to students in many developing countries. As a consequence, students perceived an incentive to obtain university degrees, even when there was a chance that they would remain unemployed for an extended period of time. When they did eventually find employment, the high wage would compensate for their earlier period of unemployment. Privately, such behaviour makes good sense in response to existing incentives; socially, however, it represents a waste of valuable and scarce resources.

The role of the international economy. In the modern view of development, an open, expanding international economy is the greatest support that the developed countries can provide for developing countries. Foreign aid can be extremely helpful in situations in which policies are conducive to development, but development will in any event be accelerated if the international economy is experiencing healthy growth. Removal of the trade barriers that developed countries have erected against developing countries is at least as important as economic aid. Trade barriers are many. They include restrictions on temperate-zone agricultural products and sugar; restrictions on the simpler labour-intensive manufactured goods (which often can be produced more cheaply in developing countries) including especially the Multifibre Arrangement under which imports of textiles and clothing into developed countries are greatly restricted; and tariff escalation, or higher rates of duties on processed products as compared with raw materials, which discourages the growth of processing industries in the developing countries. The removal of these trade barriers can help those developing countries that have already shown their capacity to take advantage of the available external economic opportunities to grow even more satisfactorily and can also provide additional incentives for other developing countries to alter their economic policies.

The Multifibre Arrangement

Population growth. Still another lesson is the desirability of slowing down the rapid population growth that characterizes most developing countries. Their average rate of population growth is about 2.2 percent per year, but there are some countries where population growth is 3 percent or more. If the aim of economic development is to raise the level of per capita incomes, it is obvious that this can be achieved both by increasing the rate of growth of total output and by reducing the rate of growth of population. Development economists of the 1950s tended to neglect population-control policies. They were partly seduced by theories of dramatically raising total output through crash investment programs and partly by the belief that population growth could be controlled only slowly, through gradual changes in social attitudes and values. But it is now recognized that some births in developing countries are unwanted. Great technical advances in methods of birth control about the same time made possible mass dissemination at very low cost. Countries where these methods were made available experienced significant declines in birth rates, although significant changes in social attitudes and values are necessary before average family size declines enough to halt population growth. As soon as birth rates stop rising, the relative increase in population in the working-age groups and the higher income available to existing family members immediately start to release resources for increasing consumption and saving.

Development of domestic industry. The positive case for the expansion of the manufacturing sector may now be considered. It is based on the general assumption that the manufacturing sector will in due course become the leading sector, drawing in workers (in part, siphoning off a portion of the increase in the labour force that would otherwise tend to drive down labour productivity in agriculture) from the traditional agricultural sector and providing them with higher-productivity jobs than could be obtained in agriculture. Agricultural productivity would necessarily be rising simultaneously, as investments in that

sector permitted increasing output. Whereas it was earlier thought that this process would follow the historical experience of countries such as England and Japan, the lesson from the successful developing countries is that by providing incentives and infrastructural support to encourage exports, there are significant opportunities for expansion of manufacturing of labour-intensive commodities, opportunities that can promote rapid growth.

Thus, given the much greater size of the international economy, and the much lower transport and communications costs that confront contemporary developing countries as contrasted with conditions in the 19th century, the potential for rapid growth is much greater now. Countries such as South Korea and Taiwan have experienced in a decade proportionate increases in per capita incomes that it took England and Japan a century to achieve. Whether other developing countries can follow this lead depends on a number of factors, including their economic policies and the continued growth of the international economy. (H.My./A.O.K.)

The central problem of countries with low per capita output is that they have not as yet succeeded in making use of their potential economic opportunities. To do so, they must achieve an efficient allocation of the available resources and provide incentives for resource accumulation. But efficient allocation of resources is not merely a matter of the formal optimum conditions of economic theory. It requires the building up of an effective institutional and organizational framework to carry out the allocation of resources. In the private sector this requires the development of a well-articulated market system that embraces the markets for final products and the markets for factors of production. In the public sector the development of the organizational framework requires improvements in the administrative machinery of the government, especially in its fiscal machinery.

Long-range planning of resource allocation

In the setting of the developing countries, one is concerned not only with the once for all problem of efficient allocation of resources but also with improving the capacity of these countries to make a more effective use of their resources over a period of time. That is to say, one is concerned not only with the static problem of the efficient allocation of given resources with the given organizational framework but also with dynamic problems of improving the capability of this framework. From this point of view, there is no conflict, as some have maintained, between the static, or the short-run, considerations and the dynamic, or long-run, considerations. The two sets of requirements move in the same direction.

The problem of the efficient allocation of investable funds in the developing countries may be taken as an example. Static rules would require the developing countries to have higher rates of interest to reflect their greater capital scarcity. But many developing countries, under the influence of dynamic theories of economic development, have used a variety of direct and indirect controls to divert large sums of capital to the manufacturing sector in the form of loans at interest rates well below the level required to equate the demand and supply of capital funds. This practice has resulted not only in a wasteful use of scarce capital resources but also in a retardation of the development of a domestic capital market. Instead of developing a unified capital market for the whole country, it aggravates the financial dualism characterized by low rates of interest in the modern sector and high rates in the traditional sector. The policy of keeping the official rate of interest below the equilibrium rate of interest also results in an excess demand for loans, leading to domestic inflation and pressure on the balance of payments and to a discouragement of the growth of domestic savings. Few private individuals are prepared to buy government securities when they frequently carry rates of interest below the rate of depreciation in the value of money. Through the pursuit of "cheap money" policies that contradict the real facts of capital scarcity, the governments of developing countries have failed to make use of the opportunity of building up a domestic capital market based on an expanding volume of transactions in government securities. (H.My.)

Developing countries and debt. After World War II it was thought that developing countries would require foreign aid in their early stages of development. This aid would supplement the capital created by domestic savings, permitting a higher rate of investment and thus stimulating growth. It was expected that their reliance on official sources of additional capital would continue until their economies had progressed enough to gain them access to private international capital markets.

Until the 1980s this pattern seemed to evolve as predicted. In the 1950s almost all capital flows to developing countries were from official sources, in the form of foreign aid from developed countries or of resources from the multilateral institutions, the World Bank and the International Monetary Fund. In the 1960s some of the export-oriented, rapidly growing countries began to rely on private international capital markets. Some, such as Singapore, attracted direct private foreign investment; others, such as South Korea, relied more on borrowing from commercial banks. In the 1970s many oil-importing developing countries were able to turn to borrowing from private sources when their economies were hit by the severe oil price increase of 1973.

The impact of oil prices

The borrowing by rapidly growing countries was of the type earlier envisaged. Investment yielded a very high rate of return in these countries, so additional foreign resources could be attracted and productively used. However, some other countries borrowed in order to offset higher oil prices and in order to maintain an excess of expenditures over consumption, without developing the highly profitable investments with which to finance the debt-servicing obligations they incurred. Balance-of-payments crises and debt-servicing difficulties had been experienced by a few countries in most years since the 1950s, but with the second oil price increase and the worldwide recession of the early 1980s, developing countries increased their borrowing and total indebtedness sharply until commercial banks virtually ceased voluntary lending after Mexico experienced difficulty meeting its obligations in 1982. The result was that a large number of developing countries were unable to meet their debt obligations, as export earnings declined owing to the recession, interest rates were rising, and new money was not forthcoming.

For many heavily indebted developing countries, the consequence was a prolonged period of slow growth or even declines in outputs and incomes. The lessons were several: The buoyant conditions of the 1970s were not likely to recur, and policies that had sustained satisfactory growth rates in those conditions were not likely to do so in the future; countries that had not yet moved away from import-substitution policies and direct governmental controls would need to undertake structural adjustments rather rapidly in order to resume their growth and to restore creditworthiness; and future private lending to developing countries would need to be somewhat more discriminating as to the economic prospects of recipient countries.

DEVELOPMENT IN A BROADER PERSPECTIVE

Modern economic development started in Great Britain, which in the 1780s accounted for a little over 1 percent of the total world population at that time. Since then, economic development has spread in widening circles to other parts of the world, spurred on by a series of technological innovations, particularly in the form of improvements in transport and communications. In the early decades of the 19th century the circle of the developed countries was limited to western Europe. By the late 19th century the circle had widened to include North America, Australia and New Zealand, and Japan. By the early 1970s about 34 percent of the total world population belonged to the developed countries, which among them had 87.5 percent of the total world GNP. What are the prospects of the still-to-develop countries of Asia, Latin America, and Africa joining this circle of economic development?

On the negative side there are a number of factors that add to their difficulties. First, the level of per capita product in the present-day developing countries is much lower than in the developed countries in their preindustrialization phase (with the exception of Japan). Second, the present-day developing countries have large popula-

tion bases and are handicapped by much faster rates of population growth. Third, they have generally a much weaker social and political framework to cope with the more explosive forces of discontent engendered by their reaction against their colonial past and by their internal economic disparities.

Factors favouring economic development

On the positive side, the present-day developing countries can draw upon a greater store of scientific and technical knowledge from the developed countries. The potential opportunities to exploit the "technological gap" are not confined to manufacturing. Modern science and technology can make immense contributions to agriculture, as illustrated by the Green Revolution created by the introduction of improved seeds and fertilizers in some Asian and Latin-American countries. Modern methods of birth control can make a decisive contribution in the race for raising per capita incomes. In addition, as the circle of the developed countries widens, they are bound to exert an increasing upward pull on the developing countries.

The economic growth of the developed countries has generally resulted in an expanding demand for the products and sometimes for direct labour services from the developing countries. But there are also the stronger localized pulls, such as the pull of the United States economy on Mexico and the pull of western Europe on the developing countries of southern Europe. The spectacular economic growth of Japan since World War II may also exert a similar pull on neighbouring countries in East Asia.

Countries such as South Korea, Taiwan, and Singapore are rapidly approaching developed-country status, and the circle is widening still farther. Rapid growth rates are being experienced by many countries in Southeast Asia. If one considers the successful developing countries of the 1950s and 1960s, it is evident that the rapid growth of the international economy was a very positive contributing factor in their success. Future widening of the circle will no doubt depend in large part on whether the growth of the international economy attains a satisfactory level.

In conclusion, the experience of the postwar years has provided many lessons that form a basis for optimism. A great deal has been learned about the types of economic policies that are conducive to rapid economic development. Rates of growth of per capita income experienced by the developing countries have been significantly higher than had been achieved by the first countries to develop. Attainable rates of growth of per capita income appear to be far above what formerly was thought feasible. The chief potential obstacles to successful development appear to be the spectre of disintegration of the international economy, should protectionist pressures be increasingly effective, and the inability or unwillingness of leaders in developing countries to adopt policies conducive to rapid economic growth. (H.My./A.O.K.)

Economic productivity

Productivity in economics is the ratio of what is produced to what is required to produce it. Usually this ratio is in the form of an average, expressing the total output of some category of goods divided by the total input of, say, labour or raw materials. In principle, any input can be used in the denominator of the productivity ratio. Thus, one can speak of the productivity of land, labour, capital, or subcategories of any of these factors of production. One may also speak of the productivity of a certain type of fuel or raw material or may combine inputs to determine the productivity of labour and capital together or of all factors combined. The latter type of ratio is called "total factor" or "multifactor" productivity, and changes in it over time reflect the net saving of inputs per unit of output and thus increases in productive efficiency. It is sometimes also called the residual, since it reflects that portion of the growth of output that is not explained by increases in measured inputs. The partial productivity ratios of output to single inputs reflect not only changing productive efficiency but also the substitution of one factor for another—*e.g.*, capital goods or energy for labour.

Labour is by far the most common of the factors used in measuring productivity. One reason for this is, of course, the relatively large share of labour costs in the value of most products. A second reason is that labour inputs are measured more easily than certain others, such as capital. This is especially true if by measurement one means simply counting heads and neglecting differences among workers in levels of skill and intensity of work. In addition, statistics of employment and labour-hours are often readily available, while information on other productive factors may be difficult to obtain. Although ratios of output to persons engaged in production or to labour-hours are referred to as labour productivity, the term does not imply that labour is solely responsible for changes in the ratio. Improvements in output per unit of labour may be due to increased quality and efficiency of the human factor but also to many other variables discussed later. There is special interest in labour productivity measures, however, since human beings are the end as well as a means of production.

The productivity of labour, land, and capital

The productivity of land, though it receives considerably less attention than the productivity of labour, has been of historical interest. In ancient and preindustrial times the products of the soil constituted the bulk of total output, and land productivity thus constituted the major ingredient in a people's standard of living. Soil of low productivity could, and over much of the Earth still does, mean poverty for a region's inhabitants. It is, however, no longer generally believed, as it was in past centuries, that a country's economic well-being is inevitably tied to the productive powers of the land, and the productive potential of the land itself has proved to be not fixed but greatly expandable through the use of modern agricultural methods. Moreover, industrialization, where it has taken place, has greatly reduced people's dependence on agriculture. These circumstances, together with expanding opportunities for trade, have enabled some countries to overcome in substantial degree the handicaps of a meagre agricultural endowment.

Attempts to measure the productivity of capital

The productivity of capital—plant, equipment, tools, and other physical aids—is a subject of long-standing interest to economists, though concern with its empirical aspects is of more recent origin. Improved statistical reporting and the availability of data in some industrially advanced countries, notably since World War II, have encouraged systematic efforts to measure the productivity of this factor. Compared with achievements in measuring labour productivity, however, the progress realized has been quite limited. There are considerable theoretical and practical difficulties to be overcome.

USES OF PRODUCTIVITY MEASUREMENT

Index of growth. A nation or an industry advances by using less to make more. Labour productivity is an especially sensitive indicator of this economizing process and is one of the major measures used to chart a nation's or an industry's economic advance. An overall rise in a nation's labour productivity signifies the potential availability of a larger quantity of goods and services per worker than before and, accordingly, a potential for higher real income per worker. Countries with high real wages are usually also those with high labour productivity, while those with low real wages are generally low in productivity. If, for the moment, other productive factors are neglected, one can see that the wage level will then be equal to the total national product divided by the number of workers; that is, it will be equal to the level of labour productivity.

The change in a nation's overall labour productivity during any given interval represents the sum of changes in the major economic sectors and industries. Some sectors and industries move ahead more rapidly than the overall average while others may gain more slowly or even decline. In the movement of a country from a level of low productivity and low income to one of high productivity and high income a strategic role is played by the industrial, rather than by the agricultural and other, sectors. In the late 18th and early 19th centuries the effect of the Industrial Revolution was felt first in the manufacture of woolen and cotton textiles, power generation, the metal trades, and machine-making industries. Along with the development of new processes came the development of

new products and services that formed the basis for new industries. An outstanding feature of these changes was an increased labour productivity that in turn laid the foundations for an enormous expansion of output. Technological change exerted its influence irregularly and unevenly and continues to do so.

In the compilation of overall averages this diversity is concealed because high rates in some industries offset low rates in others. Thus, the rate of increase of productivity for the economy as a whole varies within narrower limits than the spread of rates among individual industries would suggest. Aside from erratic short-term movements, the rate of growth of productivity may appear to be fairly stable over extended periods. A surge of labour-saving innovations would cause the overall average rate to move higher, while a technological lull would depress the average rate. History suggests that the surges tend to be associated with basic technological changes such as, for example, the steam engine, the gasoline engine, the electric motor, and the concept of the standardization of parts. Once introduced, such inventions or developments are used in many different industries. These surges tend also to be associated with such developments as, for instance, employment of the open-hearth furnace in steel manufacture or the introduction of the steam railroad.

Technological change

Productivity is valuable also as an indicator of comparative rates of change among industries and products. Growth in general can be better understood if the relative contributions of individual industries and the circumstances underlying productivity changes in each of these industries are understood.

Measure of efficiency. Productivity is also used to measure efficiency, as an aid in economic planning and forecasting, and as a means of assessing the uses to which resources are being put. As to the first of these, the efficiency of industrial operations, for instance, may be evaluated by the yardstick of output per worker or machine, and such a yardstick may also provide the basis for supplemental or premium payments for workers. When pay is based on piecework alone, labour productivity becomes the sole determinant. Productivity may also serve as a standard for grading and evaluating any group of workers performing common tasks, distinguishing the more from the less productive. And applied to equipment, productivity standards can indicate when a machine is performing poorly and is in need of service. In forecasting, productivity estimates are useful when it is necessary to be able to project the performance of the economy at some future date, given the probable size of the working force. A variant of this is common in planning for developing countries that want to increase their productivity; information about target levels of productivity, together with expectations as to the growth of the labour force and some understanding of the relation between capital per worker and output per worker, helps in estimating the amount of capital investment needed to reach the target. Again, estimates of the probable annual gain in labour productivity together with estimates of the probable annual increase in output allow one to estimate how many jobs will become available at some time in the future. Finally, productivity is a helpful analytical tool in studying the possible allocation of resources among different uses. The extent to which resources flow to various uses depends, among other things, on their productivity in each of those uses. Changes in productivity in the course of time alter the pattern of use and cause the quantities of resources required in particular uses to change. The resulting trends depend on several things. On the one hand, an increase in the productivity of, for instance, labour, since it means a decrease in labour requirements per unit of output, will tend to reduce the demand for labour. But it will also imply a cheapening of labour relative to the cost of other competing factors of production. Hence there will be a tendency to substitute labour for other factors. When labour cost represents a large fraction of total cost, a productivity increase will contribute toward a reduction in the price of the product, thereby expanding sales and with them the demand for labour. The net result will depend upon the sum total of all of these separate effects. It is by no means uncommon to find that the expansionary effects predominate, and many economists consider this to be the normal outcome. In any event, the productivity concept and data on productivity trends can contribute to an understanding of resource and output flows.

Wage and price analysis. Real average labour compensation has increased over the long run at about the same pace as labour productivity. The association of these two variables must be close so long as the labour share of total cost does not change much. If nominal average earnings were to increase more than labour productivity, labour cost per unit of output would rise and so would prices unless profit margins were reduced to compensate. In general, prices rise by less than wage rates and other input prices to the extent that total productivity rises. Productivity growth is thus an anti-inflationary factor, although inflation is basically a monetary phenomenon.

There is a significant negative correlation between relative industry changes in productivity and in prices—when productivity rises, price tends to fall. In the industrial sector of an economy in which there is a significant price elasticity of demand (*i.e.*, where price is relatively responsive to changes in demand), there is also a significant positive correlation between relative industry changes in productivity and in output—when productivity rises, output tends to rise as well. This is an interactive relationship, since the tendency of price to fall as productivity increases is reinforced by the tendency of economies of scale made possible by increased output to further enhance productivity.

Relation of prices and productivity

In dynamic economies the supply of capital has risen faster than the size of the labour force, and wage rates have risen faster than the price of capital. As a result there has been a marked tendency to substitute capital for labour in almost all industries. Yet there has been no long-term trend toward increased unemployment because real aggregate demand has tended to rise enough to absorb the growth of the labour force. Cyclical fluctuations in output and employment in capitalist countries are not the result of technological displacements of labour but rather reflect macroeconomic variables, such as growth of the money supply, that affect aggregate demand.

FACTORS THAT DETERMINE PRODUCTIVITY LEVELS

The level of productivity in a country, industry, or enterprise is determined by a number of factors. These include the available supplies of labour, land, raw materials, capital facilities, and mechanical aids of various kinds. Included also are the education and skills of the labour force; the level of technology; methods of organizing production; the energy and enterprise of managers and workers; and a range of social, psychological, and cultural factors that underlie and condition economic attitudes and behaviour.

These variables interact and mutually condition one another in determining productivity levels and their changes. Thus, in any country one expects the level of technology, the skills of the work force, the quantity of capital, and the capacity for rational economic organization to be positively correlated. A country with low productivity is likely to have deficiencies on all counts; a country with high productivity is likely to score high on all. To put it differently, the numerous productivity-determining factors behave as variables in a system of simultaneous equations, with all acting concurrently to shape the outcome. Within this system, there are no grounds for assigning causal priority to one or a few variables. All interact mutually to determine the outcome. Within certain problem frameworks, however, it may be entirely appropriate and indeed essential for explanatory purposes to emphasize certain variables over others.

Two broad problem frameworks may be distinguished, both of them of concern to students of productivity and growth. One of these involves changes in productivity over time; the other involves differences in productivity levels among enterprises, industries, and countries at a given time. Within these frameworks are many problems and subproblems, each of which may lead to a different selection and emphasis of variables.

Explanations of long-term productivity changes in a country, region, or industry usually stress technological

Changes in long-term productivity

change and, as an adjunct, changes in the quality and quantity of capital. Other variables are regarded as playing a passive role and are subordinate. The justification for this is that change in technological knowledge and the capital embodying it is both essential to substantial gains in productivity and the factor most immediately associated with those gains. It ordinarily is perceived as the leading and moving force in the process. When technological change occurs, the quality of capital improves and the amount available to aid each worker usually increases. The kinds of raw materials used may change, with better grades being required or the use of lower grades becoming possible. Changes occur in the way productive factors are organized and production is carried on. Although in some periods and in some circumstances work may have become harder and more tedious following technological advance and although the transition from land to factory has often entailed special hardships, the dominant trend has been toward shorter hours and a diminution of the arduousness of labour.

Emphasis on technological change and capital accumulation as primary forces arises also from a recognition that they are essential and unique to large and systematic advances in productivity. Those gains that can be obtained solely through a reorganization of work or the use of better raw materials or the breakdown of restraining attitudes or practices may occasionally be dramatic, but they are always limited. By contrast, very substantial gains can follow in the wake of growing technological knowledge and increasing supplies of capital. If allowance is made simply for adaptive changes in other factors, the prospects for advance become almost unlimited. Only these two factors can fairly be singled out as constituting the engines of productivity growth.

It has been noted that both the quantity of capital and its quality change as productivity increases, and it is not possible adequately to separate the two in terms of their effects. Increases in capital per worker through the accumulation of more and more of the same kinds of equipment and tools would not lead continuously to proportionate or more than proportionate increases in output per worker. They would, after a point, lead to diminishing increases and eventually even to a decline in output per worker. The onset of a decline would be far distant in an industry or economy possessed of a high level of technical knowledge but starting near the bottom of the accumulation ladder and affected by an acute scarcity of capital instruments. But an ultimate decline would be expected.

Qualitative changes in capital, reflecting advances in knowledge and skill and leading to the design and construction of improved capital instruments, offer an escape from this principle. If capital can be steadily improved over time, its expansion need not entail diminishing returns. In countries for which data from broad sectors and many individual industries are available, there is a rough correlation between growth in the quantity of capital per worker and increases in labour productivity.

MEASUREMENT OF PRODUCTIVITY

As a prelude to an examination of productivity trends over time, this section considers various methods of measuring the output and input components of productivity ratios and some of the difficulties and limitations of the resulting estimates.

Output. With respect to output, ideally the numbers of units of each category of tangible commodity or service should be counted in successive time periods and aggregated for the firm, industry, or total economy in terms of some indicator of relative importance, usually price or cost per unit as of a particular period. The unit value "weights"—price, cost, or other—must be held constant for two or more periods being compared so that changes in aggregate output reflect changes in physical volumes rather than in prices. An alternative procedure that produces the same results with ideal data is to "deflate" current values of the various items produced by index numbers that reflect relative price changes in order to eliminate the effects of price changes. Price deflation is usually employed to obtain estimates of real gross product by sector and industry to be used as numerators of productivity ratios. For tangible industrial production measures, quantities of the various commodities are generally weighted together by constant unit values.

Unfortunately, in most countries data on quantities and prices for many outputs of the finance and service industries are deficient. In the broader real gross product estimates, changes in outputs of a portion of such services are approximated by estimating changes in inputs. Estimates so derived are not suitable for productivity measurement, however. They impart a downward bias to estimates of real product and productivity for the services sector and its affected components and hence for the economy as a whole.

Output of services

Other problems in estimating output arise in adjusting estimates of outputs to take account of quality change, measuring quantities or prices of nonstandard custom-made products, and estimating outputs of nonmarket goods and services. Partial adjustments for quality changes may be made when increases in real costs per unit are associated with the improvements. But it is generally agreed that physical-volume or real-product measures fail to capture at least part of the improvements in product quality, as distinguished from relative shifts among alternative qualities (price-lines) of a given product. Methods of estimating changes in the physical volume of custom-built products, such as buildings or other major structures, have improved in recent years. But changes in the output of nonmarket goods and services, such as those of governments, households, and nonprofit institutions, must generally be measured by changes in inputs. In consequence, productivity estimates are usually confined to the predominant business (enterprise) sector of an economy.

Inputs. Labour input is relatively easy to measure if one is content to count heads of persons engaged in production or, preferably, hours worked. But in fact, the available hours data often relate to hours paid for, rather than hours worked, and these tend to rise in relation to hours at the workplace as the number of paid holidays and leaves are increased. Official estimates generally do not differentiate among various categories of labour. But some academic economists measure labour inputs by occupation and/or industry and possibly other categories and weight the aggregate in each category by a measure of the average compensation in some designated base period. As the average levels of education, training, skills, and experience of workers increase, the weighted measures rise relative to unweighted measures of labour input. Change in the ratio of the two indicates change in the quality of labour input, which is an important part of the explanation of change in productivity.

Capital input is usually assumed to change in the same direction as and proportionally to changes in the real stocks of structures, equipment, inventories, and natural resources. The rates of return on those capital goods in some base period are taken to be indicative of their productivity for the purpose of weighting them together with other factor inputs. Some analysts adjust the capital estimates to take into account changing rates of utilization of capacity; otherwise, changes in utilization rates are reflected in the productivity estimates.

Interindustry purchases and sales of intermediate products—those materials, energy, and other services that are consumed in the production process—are accounted for on a value-added basis and cancel out in the national income and product estimates by industry (one industry's output being the next one's input). But if intermediate purchases are counted as an input for comparisons with gross output estimates, they are measured in the same manner as described for outputs.

HISTORICAL TRENDS

Early industrialization. For most of humanity's history, advances in technology, productivity, and real income per capita came very slowly and sporadically. But with the development of modern science in the 17th century and the quickening of technological innovation that it sparked, the stage was set for significant improvements in productivity. The gains remained modest until the latter part of the

19th century. For the first 50 years after the beginnings of the Industrial Revolution in Britain around 1760, labour productivity grew at an average annual rate of around 0.5 percent, but it then accelerated to more than 1 percent in the 19th century. In the United States it increased at an average rate of 0.5 percent until after the Civil War.

By the latter part of the 19th century the countries of western Europe, the United States, and Japan enjoyed a marked and sustained rate of improvement in productivity generally exceeding that of Britain, the earlier leader. Growth of real gross domestic product (GDP) per hour worked in the western European countries and Japan averaged 1.6 percent from 1870 to 1950, while growth in the United States averaged 2 percent from 1870 to 1913 and almost 2.5 percent from 1913 to 1950. (See Table 1). Data for 10 additional industrialized countries indicated that much the same range of productivity growth rates prevailed for the smaller western European countries and for Canada and Australia. But much of the rest of the world had not yet begun to experience sustained growth of productivity and real per capita income.

Growth of GDP

Table 1: Phases of Growth in Labour Productivity
(real gross domestic product per hour worked; average annual compound growth rates)

	1870–1913	1913–50	1950–73	1973–84
United States	2.0	2.4	2.5	1.0
Five-country average	1.6	1.6	5.3	2.8
France	1.7	2.0	5.1	3.4
Germany	1.9	1.0	6.0	3.0
Japan	1.8	1.7	7.7	3.2
The Netherlands	1.2	1.7	4.4	1.9
United Kingdom	1.2	1.6	3.2	2.4

Source: Angus Maddison, "Growth and Slowdown in Advanced Capitalist Economies: Techniques of Quantitative Assessment," *Journal of Economic Literature*, Vol. 25, p. 65, Table 2 (June 1987).

Two percent per year may not seem an impressive number, but when compounded over a century it results in more than a sevenfold increase. The sustained and significant increases in productivity of industrialized countries beginning in the latter part of the 19th century were one of the most momentous developments in modern history, and it became much more widely diffused in later decades.

Why did the acceleration begin in the late 19th century? The great improvements in transportation and communications that were made possible by the inventions of the steam and internal-combustion engines and the telephone and wireless communications led to a major expansion of trade, both domestic and international. The British example of free trade led to some liberalization by other countries. By the turn of the century, an increasing number of large companies were beginning to conduct purposeful programs of research and development so that invention and innovation became commonplace and even expected. Educational levels rose, and business schools were founded to teach the new science of management. The growth of per capita income itself tended to raise saving rates, and investment in new plants, equipment, and natural resource development rose substantially. Finally, the growth of productivity in agriculture and increased labour mobility made possible the enormous expansion of manufacturing and, later, the service industries.

Growth of productivity in countries other than the United States accelerated greatly after World War II. The five-country average rate of growth in labour productivity (Table 1) more than tripled in the 1950–73 period compared with the preceding 80 years. After 1973 productivity growth fell by almost half in the five countries, on average, but remained well above the earlier rate. The deceleration was greater in the United States.

Before trying to explain these trends, see Table 2, which summarizes productivity changes from 1950 through a more recent year for a larger number of industrialized countries, and then see Table 4, which shows estimates for groups of countries composing most of the world.

In the 12 countries other than the United States shown in Table 2, real GDP per employed person grew between

Table 2: Real Gross Domestic Product per Employed Person
(based on own country price weights; average annual percent changes)

	1950–86	1950–73	1973–79	1979–86
United States	1.4	2.0	0.2	0.8
12-country average	3.4	4.3	2.2	1.8
Canada	2.0	2.5	1.4	1.0
Japan	5.8	7.5	2.8	2.8
Korea	5.4	5.8	5.7	4.7
Belgium	3.0	3.5	2.2	1.8
Denmark	2.6	3.7	1.4	1.4
France	3.8	4.7	2.5	1.9
Germany	4.1	5.1	2.8	1.6
Italy	4.3	5.7	1.6	1.6
The Netherlands	2.6	3.7	1.5	−0.1
Norway	3.1	3.5	2.7	2.1
Sweden	2.3	3.6	0.7	1.5
United Kingdom	2.1	2.6	1.3	1.7

Source: Bureau of Labor Statistics, U.S. Department of Labor, unpublished tabulations dated August 1987.

1950 and 1973 at an average rate of about 4 percent, about double the rate for the United States. From 1973 to 1979 the average rate decelerated to 2.2 percent a year for the 12 industrialized nations and to virtually zero in the United States. But after 1979 (and especially after 1981) the U.S. rate accelerated significantly, while the 12-nation average rate fell further to 1.8 percent, which was nevertheless still well above the U.S. rate of 0.8 percent a year.

During the entire period after 1950 there was a significant convergence of rates of productivity growth among the industrialized nations, as shown in Table 3. The average real GDP per person for the 11 countries rose from about 44 percent of that in the United States in 1950 to almost 80 percent in 1986. Furthermore, there is a significant negative correlation between the 1950 levels and the 1950–86 rates of productivity growth—those countries that started farthest behind grew most rapidly in productivity. There had already been some tendency toward convergence among the industrialized nations before 1950, but it was much stronger during the golden quarter-century following World War II.

Convergence of growth rates

Table 3: Real Gross Domestic Product per Employed Person
(based on purchasing-power-parity exchange rates; United States = 100.0)

	1950	1960	1970	1980	1986
United States	100.0	100.0	100.0	100.0	100.0
11-country average	44.3	51.7	63.6	76.2	78.9
Canada	76.9	80.1	84.1	92.8	95.0
Japan	15.2	23.3	45.7	62.7	68.9
Belgium	46.9	50.3	62.2	79.7	81.3
Denmark	49.0*	53.5	60.1	66.6	68.8
France	36.9	46.1	61.9	80.2	84.3
West Germany	32.2	49.2	61.7	77.4	80.9
Italy	30.9	43.9	66.4	81.0	82.9
The Netherlands	56.7	64.2	78.0	90.7	86.3
Norway	44.5	52.0	58.5	75.1	80.2
Sweden	44.0*	51.8	62.6	66.6	68.8
United Kingdom	53.8	54.2	57.9	65.8	70.4

*Extrapolated by author.
Source: Bureau of Labor Statistics, U.S. Department of Labor, unpublished tables dated August 1987.

Of even wider importance, most nations outside the original industrialized group also began to record substantial increases in labour productivity beginning around 1950 (see Table 4). What fragmentary information is available indicates that generally low rates of productivity growth were the norm in those countries before 1950. So World War II was a true watershed, in that after the immediate postwar period of reconstruction, most nations were able to accelerate their productivity gains markedly. Those nations, constituting all but about 5 percent of the world's population, could entertain the prospect of attaining, by or before the close of the 21st century, a standard of living comparable to that in industrial Europe in the 1980s. This prognostication assumes that productivity will continue to grow at least at post-1970 rates.

Table 4: Real Gross Domestic Product per Economically Active Person (1950, 1970, and 1980 in 1975 international dollars and average annual percentage rates of change)

	1950	1970	1980	1950–70	1970–80
Developing market economies	1,176	2,252	3,004	3.3	2.9
Low-income countries	566	813	880	1.8	0.8
Middle-income countries	1,296	2,539	3,432	3.4	3.1
Oil exporters	1,338	3,120	4,341	4.3	3.4
Relatively industrialized	2,347	4,765	6,839	3.6	3.7
Other	981	1,691	2,094	2.8	2.2
Industrial countries	5,951	10,590	13,723	2.9	2.6
Centrally planned economies	1,422	2,935	3,488	3.7	1.7
World	2,327	4,383	5,493	3.2	2.3

Source: Irving B. Kravis and Robert E. Lipsey, "The Diffusion of Economic Growth in the World Economy, 1950–80," *International Comparisons of Productivity and Causes of the Slowdown*, ed. by John W. Kendrick (Cambridge, Mass., Ballinger Publishing Co., 1984), Table A-A3, p. 145. The 1980 estimates were made from Table 3-A1 assuming that labour force participation ratios in 1980 were the same as in 1975.

The country data underlying Table 4 do not indicate a worldwide convergence of productivity levels, although some tendency toward convergence within the several groups is evident. Note that the group of low-income countries had the lowest rates of productivity advance, while the oil exporters and relatively industrialized middle-income countries had the highest rates. Whereas the centrally planned economies had above-average rates of productivity growth in the period 1950–70, after 1970 they fell below average.

The postwar growth surge. The virtually worldwide upsurge of productivity growth after World War II reflects in an important way the increasingly internationalist thinking and policy-making of leaders of the developed nations. The creation of the World Bank and the International Monetary Fund and of the United Nations and associated agencies encouraged and nurtured cooperative international economic and financial relationships. Although an outgrowth of the Cold War, the Marshall Plan unleashed a major effort on the part of the United States to aid in the reconstruction and economic development of the non-Communist world. Part of the plan called for the creation of productivity centres in member countries, which sent productivity teams to the United States to study and facilitate the transfer of advanced technology. Private lending abroad was encouraged in addition to that of the World Bank and other international lending institutions. Regional trade associations were formed to reduce trade barriers among member countries, and liberalization of international trade was promoted more generally by the General Agreement on Tariffs and Trade (GATT). As a result, world trade grew even faster than production, and most significantly it included transfers of advanced machinery and other producers' goods from the United States and other industrialized countries, which helped raise productivity of the purchasing countries.

Effect of the Marshall Plan

Multinational corporations, typically based in the United States, diffused capital and managerial and technical know-how and helped train nationals of their host countries for jobs, often including upper-level positions. International licensing of patents also helped diffuse technology. An increasing proportion of students in U.S. universities, particularly in business and engineering, came from developing countries. International professional associations and journals also aided in the diffusion of knowledge.

An important reason for the narrowing of the productivity gap between the United States and other industrialized nations after 1950 was the differential rates of saving, investment, and growth of capital per worker. In Japan the ratio of gross saving to GDP was nearly one-third, double that in the United States, and in western Europe it averaged nearly one-fourth (due in part to favourable tax laws). This higher rate of saving, creating capital for both private and public investing, was associated with a rapid decline in the average age of structures and equipment in those countries until 1973. The growth of domestic and foreign trade opened up more opportunities for achieving economies of scale in those countries as well. They also benefited more from resource reallocations, particularly the shift of labour out of agriculture and self-employment where the rates of return were lower.

After 1960 the achievement of technological parity with the United States in the ways noted above became the most important factor promoting productivity advance in the other industrial nations and in an increasing number of advanced developing countries. But, as other nations continued to approach the U.S. level of real product per person, there would tend to be greater convergence in levels and rates of growth of productivity. This would be so because innovations requiring those countries to invest in their own research and development would be more costly than technology transferred from abroad.

The slowdown in productivity growth after 1973 was almost universal. The oil-price shocks of 1973 and 1979 contributed to accelerating inflation in most countries, reducing economic profits and the rate of saving and investment. Some energy-intensive equipment was rendered obsolete. The growth of real research and development expenditures slowed, as did the pace of technological innovation. The beneficial effects of interindustry shifts of resources became less marked. The changing age–sex mix of the labour force tended to reduce productivity growth in the short run, especially in North America. And government regulations to protect the environment and promote health and safety proliferated in the '70s, increasing costs and inputs but not output as it was usually measured.

Decline in productivity growth

The reversal in the 1980s of most of those negative factors helped to accelerate productivity growth in the United States. The continued deceleration in other industrialized countries noted above probably reflected a decline in technology transfer from abroad. There appeared to be no reason, however, why the advance of productivity in the developing countries with adequate absorptive capacity might not continue for years to come. (Ma.F./J.W.K.)

Economic planning

The essential idea behind economic planning is that certain key economic decisions should be made, or at least influenced, by some central authority rather than being left to the free play of market forces. By the late 1960s the majority of the world's countries conducted their economic affairs within the framework of a national economic plan. But in the 1980s the theory and practice of economic planning went through something of a crisis. In the developed market economies the rate of economic growth slowed from the very high levels reached in the 1960s and '70s and unemployment rose significantly. At the same time, public confidence in the ability of governments to influence for the better the performance of the economy diminished. As a result, the popularity of national economic plans waned and the scope left to the free play of market forces widened. In developing countries, forms of economic planning practiced earlier yielded disappointing results characterized by the growth of heavy state bureaucracies and inefficient public enterprises. In these countries also, although the role of the state remained preponderant, market forces were increasingly relied upon to improve economic performance. In the Soviet Union and its satellites, the backward state of the economy and widespread examples of waste and inefficiency led to attempts to introduce more market solutions into the process of economic planning. These attempts proved largely unsuccessful, however, and the inherent rigidity of the Soviet economic model proved an important factor in the collapse of Communism in eastern Europe and the Soviet Union itself, beginning in 1989.

THE NATURE OF ECONOMIC PLANNING

Historically, the idea of central economic planning was associated with the criticism of capitalism as a system of anarchy and greed. Marxist critics did not give much thought to how the economy would be run after capitalism had been abolished; most of them professed to see no difficulty in organizing the society that would follow. When in 1917 the new Soviet government found itself the owner of all the means of production, it had no blueprint as to what to do next. The evolution of central economic

planning in the Soviet Union was largely a pragmatic affair; methods were tried and discarded, and new ones were introduced. The decision in 1927 to undertake rapid and large-scale industrialization required the centralizing of control, since only the government could undertake the task of marshaling the productive resources of the country to achieve its ambitious aims.

In western Europe, economic planning is adapted to a diversified economic structure, a dynamic class of business managers, and a long tradition of political and economic liberty. Consequently, although planning implies an extension of the economic responsibilities and activities of the state, the mainspring of economic growth remains the private sector. Only rarely does the state intervene directly in the affairs of individual firms. Economic planning remains indirect and takes the form of collaboration between the public and the private sectors. Producers and consumers are free to adapt their activities to changes in market conditions and relative prices. In the 1980s there was a general trend for governments to sell state-owned enterprises to the public and to reduce the extent of public regulation of particular sectors, such as air transport.

Communist methods of planning after the mid-1950s entered a state of flux, and the highly centralized administrative type of planning inherited after World War II from the Soviet Union by all the newly established Communist states underwent considerable modifications. In Yugoslavia planners followed policies very different from those of the Soviet model, and differences also emerged in the practice of other eastern European countries. In the Soviet Union itself, a debate concerning the most appropriate means for implementing plans went on for some years, but, despite numerous efforts on the part of the government to reorganize the machinery of planning, the fundamental drawbacks of central economic planning were never overcome. The Soviet Union's attempts in the late 1980s to reform its planning machinery had the unintended effect of bringing down the whole structure of central economic planning, and with it the Soviet government itself. By the early 1990s the post-Communist governments of eastern Europe and of the states of the former Soviet Union had begun making a painful transition to the diversified economic structure typical of the economies of western Europe.

In the meantime, the knowledge of the Soviet-bloc countries' long-standing difficulties had given rise in many developing countries to a repugnance to Soviet planning methods, while the methods used in the developed non-Communist countries were felt to be not directly applicable, either. There was consequently no settled planning doctrine in the developing countries, and the approach of governments remained empirical. In practice, this meant that the state played a major role in setting up new industries and in modernizing agriculture, particularly in countries of recent independence. The state budget was a major source of savings, supplemented frequently by the local currency counterpart of foreign aid. But the absence of a highly qualified civil service placed limits upon the extent and efficacy of state action. Thus, in urban areas, privately owned businesses continued to supply most local consumer goods. In agriculture, peasant proprietorship or large private estates—particularly for export products—remained the general rule. (Jo.Hac./Ed.)

ECONOMIC PLANNING IN COMMUNIST COUNTRIES

Soviet improvisations

Planning in the U.S.S.R. The kind of economic planning that was practiced in the Soviet Union and in most other Communist countries until the 1990s had developed during the 1920s and '30s in the struggle to industrialize the U.S.S.R. The Bolsheviks had seized power in 1917 without any clear notion as to how an economy should be run. No guidance was to be found in the writings of Karl Marx other than the assertion that a socialist society would operate the economy for the common good, which suggested that it would create organs of economic administration to replace the market system of capitalism. In the future Communist society there would be no money, no profit motive. No wages would be necessary to stimulate effort. It would be "from each according to his ability, to each according to his needs." Economics, a science of exchange relationships or value, would wither away or be replaced by a kind of higher management. The Bolshevik leader N.I. Bukharin wrote in 1920:

> As soon as we deal with an organized national economy, all basic "problems" of political economy, such as price, value, profit, etc., simply disappear . . . , for here the economy is regulated not by the blind forces of the market and competition but by the consciously implemented plan.

The leader of the Bolsheviks, Vladimir Lenin, shared these somewhat utopian expectations. It is not clear from his pre-1917 writings just what kind of economic organization he had in mind for Russia should he achieve power. As it turned out, the Russian Revolution of October 1917 was accompanied by economic breakdown, a refusal of cooperation from officials and managers, civil war, and uncontrollable inflation. Partly under the stress of these circumstances, partly from ideology, the Bolsheviks moved to establish thoroughgoing state control over industry and trade, nationalized all economic property including land, declared all private enterprise illegal, and demanded that the peasants deliver all farm surpluses to state procurement organs. Money lost all value.

On paper, this period of War Communism, as it is now called, was one of centralized planning. All economic units, except the peasant producers, were subjected to orders from the government's Supreme Council of National Economy (V.S.N.Kh.). But this initial essay in planning was a failure—except insofar as it facilitated the concentration of the few available resources for the civil war fronts. Rationing functioned spasmodically, there was famine, and output fell drastically.

From Lenin to Stalin

The controversies of the 1920s. In 1921 Lenin introduced the New Economic Policy (NEP). Small-scale private manufacturing, private trade, and free sale of peasant surpluses became legal once again, while the state retained the "commanding heights" (*e.g.*, large-scale industry, foreign trade, banking, transport). The state sector continued to be operated under the aegis of V.S.N.Kh. by trusts and enterprises with state-appointed managers. In 1921 the State Planning Committee (Gosplan) came into existence to advise the government and its economic alter ego, the Council of Labour and Defense, but planning was still a shadowy process. Trusts and enterprises had considerable autonomy and were free to make agreements and grant credits to one another. The planners made forecasts, and government policy decisions influenced the level and direction of state investments; but there was no integrated system of production and allocation planning, even in the state sector, while the private sector was not directly planned at all. In 1924 only 35 percent of the national income, 1.5 percent of agricultural production, less than half of all retail trade, and three-quarters of industrial output were "socialized"; the rest was private.

In 1926–28 a vigorous discussion raged about the future basis of planning. Two schools of thought developed, one advocating "genetic" and the other "teleological" planning. The former, composed of the more moderate and cautious planners, believed that plans should be based on existing trends in the economy and reasonable projections thereof. The latter considered that drastic measures were necessary to speed up the industrialization of the country, and this "teleological" school produced extremely ambitious drafts of the First Five-Year Plan. The radicals conceived the plan as taking precedence over all previous economic decisions so as to enable a sharp break with the past. With the support of the Soviet leader Joseph Stalin, it was their view that won.

Priority for heavy industry

The First Five-Year Plan. The First Five-Year Plan (1928–1932), which was later said to have been carried out in four years, called for immense investments in heavy industry; for example, steel output was to be more than doubled by 1932. Amid great confusion, the planning mechanism was overhauled, and gradually a "command economy" was established. In this system, subordinate units of the economy (*e.g.*, industrial enterprises) operated in accordance with administrative instructions, and they did not decide their inputs or outputs by negotiation with other enterprises, these being determined by the planners.

The sole effective criterion of management decision became conformity to plan—*i.e.,* to the instructions issued by the central administrative planning organs. In this way, the political authorities achieved a high degree of control over resource allocation and were able to enforce their priorities. Consequently, when the First Five-Year Plan ran into trouble, the government was able to insist on the fulfillment of most of the plans for key sectors of heavy industry, at the cost of a drastic fall in availability of consumers' goods.

The structure of the Soviet planning system. A rearrangement of the planning system was the necessary consequence of the new tasks it was called upon to perform. In 1932 three People's Commissariats (for heavy, light, and timber industries) replaced the V.S.N.Kh.; these were further split, and by 1939 the industries of the U.S.S.R. were run by 21 People's Commissariats (the numbers varied in subsequent years). Each commissariat (renamed ministry in 1946) controlled a branch of industry, either directly or through a ministry in one of the union republics. The ministries issued instructions to "their" enterprises, organized the supply of materials and components, and also disposed of the output.

Control from the top

At the apex of the system stood the leaders of the Communist Party, who decided the policy objectives in economic as in other matters and who made choices as to the means of achieving those objectives. All key appointments in the economic hierarchy were made or confirmed by appropriate party committees.

The work of Gosplan. It was Gosplan's task to "translate" the politically determined objectives into a consistent set of plan targets. There had to be coherence between production and supply at all times, as well as between investment plans and the current production of capital goods. Foreign trade also had to be taken into account, as a drain on available resources (exports) and as a source of needed goods (imports). The planners proceeded by drawing up a series of material balances, which expressed anticipated supply of, and demand for, all key commodities. The successive versions of the plan were revised until a general balance was attained, since it was no use planning an increase in production of any item if the necessary additional machinery, raw material, and fuel could not be made available. The task was of special complexity in the short term (*i.e.,* within a period of a year), since the plan had to take the form of millions of consistent instructions to thousands of enterprises to produce, deliver, transport, and process millions of commodities of a great many shapes, sizes, and types.

Balancing supply and demand

Needless to say, all these decisions must be made somewhere in all economic systems. The Soviet type of "command economy" developed under Stalin, however, provided no criterion for decentralized decision making such as is provided, however imperfectly, by the market mechanism in Western capitalist countries. Consequently, the coordination of all these decisions had to be consciously achieved by the planners. In practice much depended on proposals from below, since the planners suffered from information overload. The actual plans were necessarily aggregated (*e.g.,* tons of metal, millions of square metres of cloth, millions of rubles' worth of construction or of furniture), so that decisions on the product mix were necessarily decentralized. The resultant malfunctioning came to be much criticized in the Soviet press. Quality was often sacrificed in order to fulfill the plan in quantitative terms; planned targets expressed in tons, for example, encouraged excessive weight in the product concerned, while targets expressed in rubles discouraged economy and rewarded the use of expensive materials. Plan-fulfillment as a dominant criterion of success stimulated management to conceal their productive potential so as to get an "easy" plan, while fears of supply shortages encouraged hoarding. Soviet critics increasingly pointed to the rigidity of prices, which did not reflect supply–demand conditions. The planners claimed that it was their task, not that of the price mechanism, to ensure balance between supply and demand, but the enormous complexity of their task made it impossible for them to do so.

Low growth rates in the late 1970s and early '80s, on top of continued shortages and corruption, alarmed the Soviet leadership. Many proposals were aired as to how the system might be changed. A series of reforms were in fact promulgated (notably in 1965 and 1974), but these were soon criticized as having been inconsistent and halfhearted.

The Gorbachev reform agenda. The program of reform proposed and undertaken in the period 1987–90 under the leadership of Mikhail Gorbachev represented a truly radical change in the nature of the Soviet system, the first since the early 1930s. In this program it was intended that the bulk of the product mix would be decided not by the planners but by management, in negotiation with their customers or with the wholesale-trade organs. The need for competition was explicitly recognized, both between state enterprises seeking customers and between them and newly legalized cooperatives (more or less free enterprises). Enterprises were to be protected by law against arbitrary exactions by their superiors. An end was decreed to "soft" credits and subsidies, leaving open the real possibility of bankruptcy. Large enterprises were to be allowed direct access to foreign markets.

Reforms along these lines were gradually introduced, but some formidable obstacles proved impossible to surmount. One was chronic shortage, which continued to stimulate hoarding and compelled the continuation of material allocation. Prices were only slowly and with difficulty reformed. The declared aim of speeding up growth led to the survival of growth targets, in the familiar units of rubles, tons, and square metres, although the reformers aimed to abolish such targets. The priority of centrally determined objectives was assured by the system of so-called *gos-zakazy* (state orders), and these could cover the major part of the output of many enterprises. There were, moreover, serious problems of ideology (the enhanced role of the market came into conflict with traditional Marxist views) and bureaucratic resistance. Deeper reforms that were proposed threatened the fundamental powers of the Communist Party and its officials. In the meantime, the central government watched its authority over economic decision making steadily erode at the republic and regional levels, largely owing to Gorbachev's more liberal policies. Central economic planning ceased to have any meaning as many enterprises, effectively freed from government oversight, tried to cope in an economy that as yet lacked the free play of market mechanisms. With the collapse of the Soviet central government in late 1991, economic policy-making devolved upon Russia and the other newly independent republics of the former union, most of whom appeared committed to a diversified economic structure in which central planning would play a much-reduced role.

Agricultural planning. Agricultural planning in the Soviet Union had a peculiarly difficult history. With priority given to industrialization, agriculture during the regime of Stalin was essentially treated as a source of cheap food and materials for the cities. The peasants were, in fact, expropriated by force in the period 1930–35, and the bulk of them were compelled to join collective farms (*kolkhozy*). While in Soviet ideology state farms, operated like factories with wage labour, were preferred to collective farms, they remained of relatively minor importance until after 1954. Mechanization was for many years confined to a very few crops and especially to grain growing. The entire system was primarily designed to ensure deliveries of produce at low prices, and the planners and administrators concentrated on procurements, while production plans were seldom, if ever, fulfilled. Under Nikita Khrushchev in the late 1950s and early 1960s there was a substantial change of policy, with greatly improved prices and a major investment program designed to restore agriculture to health.

Farm collectivization

This policy was continued under Leonid Brezhnev in the 1960s and '70s. Despite very large investments and higher farm prices, however, output rose slowly and costs rose quickly, necessitating very large subsidies. Peasant incomes rose, but incentives to work on the large state and collective farms were ineffective, and millions of townspeople had to be mobilized annually to help with the harvest. An important reform was the spread within state and collective farms of the use of autonomous work groups that

were paid according to results. In 1987, proposals were adopted that would allow the leasing of land to families over and above the small plots and privately owned livestock that most rural residents had and that even as late as 1986 were producing 25 percent of the Soviet Union's entire agricultural output.

As the authority of the central government crumbled in 1990–91, many state and collective farms gained de facto control over their own affairs, though few used this to any distinct advantage. More profound changes seemed likely as a result of the breakup of the Soviet Union in 1991 and would probably involve the reversion of farmlands to private ownership in some republics.

Planning in other Communist countries. In other Communist-ruled countries the Soviet system was extensively copied, even in minor details, until 1956. After that date much depended on choices made by the party leadership of each country. Both Yugoslavia (in the 1960s) and China (in the 1980s) decentralized control over major sectors of their economies and introduced individual incentives on a significant scale. The Soviet Union's satellites in eastern Europe, by contrast, maintained fairly rigid centralized controls until 1989–90. At that time, the Soviets abandoned their political-military control over the region, and most eastern European countries used the opportunity to begin moving toward a free-market economic system, however haltingly and even painfully.

Poland. Poland's unsound economic policies in the 1970s led to serious domestic imbalances and a growing foreign debt and contributed to the political-economic crisis of 1980–81. Martial law, imposed in 1981, made possible the imposition of a very sharp rise in consumer prices, and the regime then adopted a radical reform designed to greatly strengthen the market mechanism. Its implementation, however, was delayed by the chronic shortages and imbalances inherited from the previous period. It is noteworthy that the bulk of agriculture in Poland remained dominated by private peasant smallholders, who were free to sell what and when they wished. Beginning in 1990, the new post-Communist government of Poland abandoned price controls and subsidies and undertook a major currency reform in a drastic program to convert the Polish economy to a free-market basis. The privatization of the larger state-owned enterprises proceeded relatively slowly, however, as in other eastern European countries.

Czechoslovakia. Czechoslovakia's centralized economic system was in the process of being reformed in 1968, when fears of more fundamental political change brought about Soviet military intervention, which had the side effect of halting the economic reform process. Following the events of 1989–90, Czechoslovakia moved in the same general direction as Poland. State subsidies on many items were reduced, prices were decontrolled, and the private ownership of industrial and commercial enterprises and of farmland was legalized and even encouraged. Larger industrial enterprises were converted to joint-stock companies, and their shares were sold to the public.

East Germany. East Germany's industrial planning was based upon a set of monopolistic cartels (*Kombinate*), which had considerable autonomy in carrying out the tasks of satisfying the needs of domestic customers and of export markets. Perhaps because of traditional German organizational skills and work ethic, the system was more efficient in operation than those of most other countries in the Soviet bloc. It remained woefully inefficient by the standards of the free-market economies of western Europe, however, as became clear following West Germany's historic unification with East Germany in 1990. When deprived of their state subsidies, most eastern German industries proved unable to survive in free competition with those of western Germany or with other European Community countries. As a result, eastern Germany's rapidly shrinking industrial sector quickly came to depend on subsidies from the German government and on massive new plant investment by corporations based in western Germany.

Romania. Of all the Soviet-bloc nations, it was Romania that most fully retained Stalinist methods, both in the economy and in politics, into the 1970s and '80s. Unsound economic policies led to a long-lived situation of crisis and acute shortages, especially of energy and even of food. The resulting widespread deprivation sparked a popular uprising in 1989 that overthrew Romania's longtime leader, Nicolae Ceauşescu. But, as in some other eastern European nations, the end of Communist rule in Romania was followed by a sharp economic decline: the closing of unprofitable state-supported industries resulted in falling production and rising unemployment, while shortages of food and other consumer goods continued and even worsened.

Hungary. In 1968 Hungary adopted a system of market socialism that left each enterprise management very largely free to determine its own production program. The central planners were no longer to set obligatory production targets. While some prices remained controlled, others were set free. Enterprises were also given some freedom to buy and sell abroad, and efforts were made to link Hungarian prices with those on world markets. Profits became the principal measure of managerial success, and bonuses based on profits had an appreciable effect on managers' and workers' incomes. Large-scale investments were still controlled by the central planners, but the enterprises were required to finance roughly 40 percent of all investments out of their own resources.

Market socialism in Hungary

Balance of payments difficulties and internal pressures (*e.g.,* for subsidies to unsuccessful enterprises), however, led to severe strains, and output and living standards stagnated after 1978. Agriculture, however, dominated by genuinely autonomous cooperatives and a flourishing private sector, continued to do well. Hungary also had a sizable "second economy," with a variety of legal small-scale private enterprises.

Yugoslavia. The Yugoslav Communists developed their own conception of socialist planning after their break with Moscow in 1948. The collectivization of agriculture was abandoned in the early 1950s. The control of the state-owned enterprises was given to workers' councils that would decide their own production programs according to profitability, with prices subject to negotiation. Investments were partly controlled by the enterprises themselves out of profits or by the central planners, partly financed from bank credits. But lack of effective central control, and rivalries between national republics, gave rise in the 1980s to a serious economic crisis led by a rapidly rising rate of inflation. These economic difficulties foreshadowed Yugoslavia's breakup in the early 1990s.

China. Chinese Communist planning at first followed the Soviet pattern. In 1958, however, came the Great Leap Forward, an effort to speed up progress by shifting rural manpower into manufacturing. This failed disastrously, and the Chinese Communist leadership had to devise its own planning methods, adapted to a vast country with poor communications and a low stage of economic development. After the social-political cataclysm known as the Cultural Revolution and the death of Mao Zedong, reformers led by Deng Xiaoping came to power in the late 1970s and launched a major shakeup of the system. Agriculture was decollectivized, small-scale private trade and workshops were legalized, and the role of market forces was substantially increased. Larger-scale industry remained subject to central planning controls, though there, too, market-type reforms were experimented with. While there were successes, balance of payments problems and inflationary pressures continued to cause some anxiety. Agricultural output rose sharply at first, but concern over shortfalls in cereals production continued. In China, too, the search went on for the elusive optimal balance between plan and market.

Major reform movement

Assessment of Soviet-type planning. The Soviet type of planning grew up under the special conditions prevailing in the U.S.S.R. and was adapted to the task of speedy industrialization of a poor country, with strong emphasis on heavy industry, explicable partly by the logic of industrialization (steel and machinery are more conducive to industrial growth than textiles and jam), partly by concern for military potential. The system made it easy for the authorities to attain a high rate of forced saving and investment and a rapid buildup of basic industries, though at the cost of neglecting for many years the elementary needs

of the citizens. Insofar as the investment plans of most basic industries depended in the last resort on a quantitative estimate of future demand, the Soviet system was reasonably well adapted to making such estimates, since so much of the additional demand was a consequence of the planners' own decisions.

In practice, of course, Soviet-type planning was not always able to realize these potential advantages. There were repeated instances of overinvestment, followed by the abandonment or freezing of partly finished projects. Experience also showed that the separate administrative units into which a nationalized economy must be divided can take as narrow and short-term a view as any capitalist entrepreneur. Thus, the most easily accessible forests were cut, the richest sources of iron ore exhausted, and fallow land put under the plow in order to fulfill current plans, with little consideration of the consequences. A centralized system of material balances is not insurance against erroneous forecasting. The material-balances approach exercised a conservative influence, perhaps because it was simplest to plan on the assumption that the various technical coefficients would remain constant. Innovation was often resisted, and the influence of user demand was weak. It must be borne in mind, of course, that Western economies also have many imperfections. A theoretical model of centralized planning works as smoothly and as efficiently as a theoretical model of a perfectly competitive market, but neither exists in the real world.

Shortcomings of central planning

Following the death of Stalin in 1953, the Soviet economic system was presented with new problems. It showed itself unable to cope effectively with the finer adjustments required in a sophisticated industrial economy. In particular, the system was not able to stimulate the adoption of new technology despite heavy expenditure on research. It dealt very clumsily with the satisfaction of consumer demand, though this became more important in the changed political conditions, not only in the Soviet Union but also in most of its European allies. The unsuitability of the traditional Soviet planning model in a modern, highly technological, and intensely competitive world economy became painfully clear to the generation of Soviet leaders who came to power in the mid-1980s. But their efforts to incorporate with socialist planning the flexibility and grass-roots enterprise that come with market mechanisms also failed, largely because central planning had become indissolubly tied to the totalitarian structure of state power in the Soviet Union. (A.No./Ed.)

ECONOMIC PLANNING IN NON-COMMUNIST COUNTRIES

Planning in developed countries. Since the end of World War II in 1945, most non-Communist developed countries have practiced some explicit form of economic plan. Such countries include Belgium, Canada, Finland, France, Germany, Ireland, Italy, Japan, The Netherlands, New Zealand, Norway, Portugal, Spain, Sweden, and the United Kingdom. Planning as a focus for economic policy-making in these countries had its heyday in the 1960s and '70s. After that time, although the formal mechanisms for working up the national economic plan remained in existence, their impact upon national economic policy-making was much diminished. Governments harboured narrower ambitions, and public opinion came to expect less from government action.

Origins of planning. Until World War II there was no serious attempt at economic planning outside the Soviet Union. During the Great Depression of the 1930s, many governments were forced to intervene vigorously in economic affairs, but in a manner that amounted to economic warfare; this intervention took the form of giving increased protection to domestic producers against competition from abroad; of acquiescing in the formation of cartels and other arrangements among producers to raise prices and reduce competition; and of higher levels of government spending, some of it for relief and some of it for armaments.

Planning for economic warfare

At the end of the war there was a shift to the left in the politics of some of the countries, and with it a turn to more positive forms of government intervention. In Great Britain the Labour Party secured a large majority in Parliament in 1945, and with it a mandate for policies aiming at more social equality. In Scandinavia, particularly in Sweden, moderate left-wing traditions in government made a transition to planning politically acceptable. In France, left-wing groups, including the Communist Party, emerged as the dominant political force after 1945 with programs of far-reaching social reform. More important, a group of eminent public servants, engineers, and business leaders—continuing a tradition of French 19th-century capitalism known as Saint-Simonianism—were in favour of the state taking a leading role in economic affairs.

While the initial impulse to planning came from the political left, actual decisions by governments to plan were based on practical considerations rather than on political doctrine. The decision to plan most often followed a crisis in a country's economic affairs, as was the case in France after World War II, when there was an urgent need to reconstruct and modernize the economy. In the United Kingdom the setting up of a medium-term plan accompanied the emergency measures taken to deal with a balance of payments crisis in July 1961; and the Labour government's National Plan of September 1965 was formulated in similar circumstances. In Belgium and Ireland dissatisfaction with the past performance of the economy was a major reason for planning. Belgium had not shared in the European prosperity of the 1950s, and accordingly, in 1959, the government adopted a plan aimed at an increase of 4 percent a year in the GNP, practically double the rate achieved from 1955 to 1960. Its planning methods were modeled on those of France.

The French example also influenced planning in other European countries. In Great Britain a Conservative government undertook, during a balance of payments crisis in July 1961, to set up a National Economic Development Council to draft a five-year economic plan that would emphasize much more rapid economic growth. The Netherlands, which had been very successful since the war in achieving balanced economic growth, initiated five-year plans in 1963 through the medium of the Central Planning Bureau, which had for some years been advising on national budgetary policies. Italy had first turned to planning in the 1950s, when a plan for the development of southern Italy was launched; later, attempts were made to extend this example of regional economic planning into a plan for the national economy. Even in West Germany, where the Christian Democratic governments had emphasized a policy of strengthening the free market, a need for some central management of the economy was increasingly recognized.

Economic planning in the developed countries has always been pragmatic rather than inspired by an attempt to apply preconceived ideological doctrines. In the 1980s, governments in most of these countries swung to the right of the political pendulum and were therefore less sympathetic to the idea of economic planning, which therefore took a back seat in national economic policy-making. The problems that the developed countries faced (chiefly slow growth and high unemployment) were thought not to be amenable to more state action. Indeed, the cost of financing government was thought in influential circles to be stifling private initiative. In the same way, many enterprises under public ownership were "privatized" (that is, returned to private ownership), and the scope of government regulation of the economy was notably reduced. In the view of a new generation of policymakers, the major role of government in promoting economic growth was, first, to provide a stable, noninflationary framework for enterprises to make their decisions and, second, to support the emergence of the new "information society" through improved education and technical training and research and development programs.

Objectives of planning. The plans drafted in the developed countries postulate target rates of growth and some indication of the choices to be made in allocating rates of increase to the various kinds of expenditures on available goods and services (private and public consumption, social investment, directly productive investment, stocks, and exports). Further, as the plan has to be balanced, the total of these components of overall demand must fall

Growth, balance, and consistency

within the probable total available supply of goods and services, after allowing for the desired level of the current balance of payments surplus. The rates of increase of some types of demand—directly productive investment and stocks in particular—are fixed within fairly narrow limits by technological considerations once the overall rate of growth of output has been chosen. But an important area of choice concerns the respective rates of growth of private consumption by individuals and what has come to be called collective consumption; *e.g.*, education, health, urban facilities, provision for culture.

Another planning objective that has become very prominent (and that figures in the plans of Great Britain, The Netherlands, France, and Italy, to mention only a few) is the correction of imbalances in regional development.

One of the most important functions of economic planning is to achieve consistency among different economic objectives. Some desirable goals are likely to conflict with others. While it may be possible, for example, to stimulate the economy so as to obtain sharply higher levels of output and employment, the measures required to do this may also produce rapidly rising prices, which in turn will lead to rising imports and falling exports; the result may be a balance of payments crisis.

Public discussions of objectives

Stages of planning. One of the chief merits of national planning is that it publicizes the choices before a country and encourages discussion of them. In France the Parliament is consulted on the broad outline of the five-year plan, which is presented in terms of a number of alternative balanced sets of objectives. More government spending on social services, for example, can be shown as implying a slower rise in individual incomes after taxes, or—a rather more difficult choice to make explicit—a higher growth rate can be shown as requiring greater willingness on the part of producers and consumers to adapt themselves to changes in markets and technology.

In other countries the choice of objectives may be left to the government, which thus makes the plan a part of the program upon which it will stand or fall at the next election. Or the government may prefer to detach itself almost entirely from the plan. In West Germany, where planning is less explicit, projections of economic trends are set out in a technical document that does not have the formal status of a government draft and circulated to the governments of the *Länder* (states) and to employers and trade unions. Such projections are thought to have some influence on public thinking and on the expenditure plans of business and government. In Great Britain, when the Conservatives returned to power in 1970, there was less enthusiasm for public discussion of planning objectives than there was under the preceding Labour government. The Dutch procedure is to leave the main responsibility for drafting projections of economic trends to the Central Planning Bureau. In that country, however, business and labour have generally been ready to take account of such projections when drawing up their own plans. In Belgium, after the period of strained relations between the main language groups during the 1960s, regional considerations have been very much to the fore in all planning discussions.

An important issue for the European countries in the late 20th century has been the impact on their economies of the European Economic Community (Common Market) and the aim of achieving a single, unified market by 1992.

Bringing the public into the planning process

Since targets in this type of planning do not constitute orders to producers and consumers to do or not to do particular things, the plan has to take account of what private firms say they intend to do, or could do, during the period of the plan. Drafting a plan, therefore, requires arrangements for bringing the representatives of the private sector—both employers and workers—into the planning process. This is usually done by setting up a tripartite body (the High Planning Council in France, the Council for Economic Planning in Sweden, for example) where representatives of the private sector are brought together with representatives of the government.

Government departments in most countries draft programs several years in advance for expenditure on such public projects as education, road building, urban improvement, and hospital construction. But it often happens that these programs are not mutually consistent. One of the advantages of an overall plan is that the confrontation and coordination between the various programs has to be done in an explicit manner. The task can be difficult; watertight administrative compartments are not always easy to break down.

Efforts to incorporate into the plan all that can be known about the intentions of the private and public sectors need to be supplemented by more general economic analysis. As has been stressed already, the plan should be, above all, coherent; and the best way of ensuring this is to build up the planned targets within the framework of the national income accounts. Since World War II, such accounts have been drawn up on an annual basis in all the developed countries. From this it is only a step to projecting the national accounts for several years ahead.

Innovations in planning

Earlier types of economic planning leaned heavily upon the method of economic balances. This consists of setting out the quantities of economic resources that will be available during the plan period and comparing them with the quantities demanded by the plan. Four balances are of key importance: the demand for and supply of goods and services; of savings; of manpower; and of foreign exchange. The notion of balance is a valuable one in planning, since no plan can be successful if it outruns the available resources. The method has its difficulties, however, because of the numerous interactions among different sectors of the economy, with the consequence that an adjustment in one set of balances requires adjustments in the others—a complex and time-consuming process. The method of balances is also unable to throw light upon a more fundamental aspect of economic decision making, the need to choose among alternative courses of action on their own merits.

Other methods of planning that have in varying degree replaced the method of balances include mathematical model making and cost–benefit analysis. A mathematical model consists of a series of equations that describes the structure and working of the national economy, enabling various sets of targets to be "tried out" by feeding their values into a computer. It is too early, however, to claim for economic models any clear superiority over the more pragmatic method of economic balances. The most systematic use of models has been in The Netherlands and France. In the Scandinavian countries, a strong mathematical tradition among economists has favoured their adoption.

Cost–benefit analysis, sometimes known as the planning–programming–budgetary system (PPBS), represents an effort to improve the planning of government expenditures. Starting from the fact that public expenditures are not sensitive to the economic considerations of price and profitability but that they nevertheless use up scarce resources that have economic value, PPBS attempts to provide rules of calculation when deciding upon the allocation of these resources. A first step is to break down public expenditure into its main functions and to divide each of the latter into programs that can be identified with government policy objectives. Then an effort is made to evaluate the effectiveness of each program in achieving its declared objectives, together with a consideration of alternative ways of achieving the same objectives and the costs of those alternatives. The U.S. government pioneered in the application of PPBS to government activities in the 1960s. Great Britain introduced it in the Ministry of Defence in the late 1960s and then began to extend it to other departments, particularly in education and science. In France the government decided to apply the system in 1968, first in the Ministry of Defense and then in relation to energy, town planning, and such departments as posts and telegraph. By the early 1970s PPBS had become an integral tool of national economic planning.

Problems of flexibility

In their planning, all the non-Communist countries leave a large margin of initiative to individual producers and consumers; *i.e.*, they rely upon market mechanisms rather than upon direct controls. There is no contradiction between this state of affairs and the existence of a plan. First, as already noted, the fact of associating the private

sector with the drafting of the plan encourages its representatives to be dynamic in their behaviour, a prerequisite for any successful growth policy. Second, when a plan is drafted in this way, all sectors are encouraged in making their own plans to adopt the same general assumptions about the growth of the economy. In France this aspect of planning is called "generalized market research." Third, participation in drafting the plan can help to make the representatives of the employers and workers more conscious of existing obstacles to growth and encourage them to find constructive solutions for overcoming them.

But the government is not freed from participating actively in carrying out the plan. The government's budget and its monetary policy must be used in such a way that overall demand rises as steadily as possible along the desired growth path. The problem of avoiding checks to growth caused by balance of payments difficulties is likely to call for special attention.

Direct government intervention

There is no easy solution to the problem of how to reconcile five- or six-year plans with the need to adjust economic policy to the phases of the business cycle. In Japan, a country that has experienced several sharp though brief recessions since the end of the postwar reconstruction period, governments are willing to cut back their public expenditure programs and to take restrictive fiscal and monetary measures because they believe that the economy will soon rebound once the restrictive measures have done their work. Japan's remarkable economic growth has borne out this confidence. In the less flexible economies of western Europe, governments are more circumscribed in their action. France, Germany, and The Netherlands have instituted systematic reviews of their plan targets every one or two years and are quite prepared to revise the targets if necessary. In addition to these general measures for implementing plans, some governments attempt to influence more directly the behaviour of private firms by granting depreciation allowances for investments and tax exemptions for some activities, such as expenditure on research.

But it is in the field of regional development policy that governments intervene most actively to influence the decisions of the private sector of the economy. There is widespread use in western Europe of tax incentives and grants from public funds to encourage private firms to expand in less-developed regions. Frequently, permission to set up a new factory, or to expand an existing one, in already congested areas where labour shortages are acute is refused. In addition, governments endeavour to stimulate growth in the less-developed areas by improving transport facilities, housing, and urban infrastructure generally. These efforts have not shown any significant success in raising incomes or reducing unemployment in the less-developed regions. This does not mean that the regional policies have been without effect, however, for without them the regional disparities might have worsened. Present-day plans for regional development attempt to concentrate on a few selected "growth points" where it is thought that favourable geographic characteristics or the presence of raw materials or other advantages offer the most hope of success.

Assessment of planning. By the 1970s planning had become more flexible and selective than in earlier years, and the trend continued and even accelerated in the 1980s.

Economic growth as an object of policy

The general consensus was that the government should seek to create the fundamental conditions that would encourage growth; this would include measures to establish and maintain competition. The corollary was that governments should try to avoid applying detailed controls over the private sector in peacetime, since these lead to reduced efficiency.

Some critics of planning have charged that the planners put too much emphasis on measures to accelerate economic growth, overlooking the social costs involved. A difficulty with simple growth targets is that they do not measure the increase in side effects such as pollution, noise, and the destruction of nature; on the contrary, they show the expenditures on combating these effects as part of the growth itself. (For example, expenditures on conservation or smog abatement are included in the statistics of national income and GNP.) Similar contradictions are found in the easy equation of economic growth with the general welfare: it is possible for income per head of the population to rise while the incomes of certain groups fall; quite frequently, some groups, such as the aged, handicapped, unemployed, and certain ethnic groups, do not share in the increasing prosperity of their country. The general welfare obviously includes elements such as health, housing, education, and economic opportunity as well as economic growth. This concern with the qualitative aspects of economic growth has left its mark upon the objectives written into the economic plans, which increasingly spell out general social aims.

Governments have also adopted a more flexible approach to the setting of targets. During the 1960s, mainly under the influence of French practice, targets for the private sector were often spelled out in detail. But experience showed that elaborate targets were rarely achieved, although they were likely to be considered by public opinion as representing firm commitments by the government. Since then, care has been taken to distinguish between firm targets and estimates. The firm targets are set only for areas over which governments have a considerable degree of control. While the government may have some influence on output in manufacturing industry, this depends more directly upon such things as the state of business conditions abroad, the purchasing habits of consumers, trends in prices and incomes, and so on. It is noteworthy that Japan, which holds the record for economic growth since World War II, has never used detailed output targets in its multiyear plans. (Jo.Hac.)

Planning in developing countries. Since the end of World War II, it has become an accepted practice among the governments of the developing countries to publish their "development plans." These are medium-term plans, usually for a five-year period. The aim is to select a period long enough to include projects spanning a number of budget years but not so long as to delay periodic assessment of the development effort stretching over a series of plans. The development plan attempts to promote economic development in four main ways: (1) by assessing the current state of the economy and providing information about it; (2) by increasing the overall rate of investment; (3) by carrying out special types of investment designed to break bottlenecks in production in important sectors of the economy; and (4) by trying to improve the coordination between different parts of the economy. Of these, the first and fourth are perhaps the most important and the least understood function of economic planning. The other two functions of planning cannot be efficiently carried out without ample and reliable information, nor without effective economic coordination between the different government departments and agencies within the public sector and the private sector. In most developing countries, information about the economy is scarce, and planning has provided the impetus to acquire and analyze the necessary data in order to provide a better understanding of the functioning of the economy. In order to improve coordination it is necessary to spread reliable economic information to indicate the future course of the government's economic intentions and activities so that the people concerned, both in the public and the private sectors, may make appropriate plans of their own to bring them in line with the government's plan. In fact, this may be regarded as the main reason for publishing development plans, although this point is not always clearly appreciated by the governments that issue them.

Why poor countries plan

Approaches to development planning. The newly independent countries, just starting to plan their economies, usually begin with a simple type of development plan. In most cases this is merely an ad hoc list of individually conceived social and economic projects that the various government departments have submitted for the plan. So long as the projects are well selected (say, to break some obvious bottlenecks in production) and are well designed in a technical sense, such a simple plan may be quite serviceable. But it tends to suffer from a number of weaknesses arising from insufficient coordination. (1) Since the projects are drawn up on a piecemeal basis in separate

From simple to comprehensive planning

government departments, there is usually no systematic attempt to compare the relative costs and benefits of the plans proposed by the different departments on a uniform basis. As a consequence, the collection of projects included in the plan may or may not represent the most productive pattern of investing the available resources of the government. (2) A lack of coordination frequently leads to wasteful duplication and a failure to take advantage of complementary relationships between individual projects. (3) A simple listing of the projects does not provide a clear-cut system of priorities in their implementation. Typically, the projects that are relatively easy to implement are pushed far ahead of others that, although requiring a longer time to prepare and implement, may have the potential to contribute more directly to the expansion of national output and government revenue. This can have serious budgetary consequences when the projects that are easier to implement generally happen to be in the field of social welfare, education, and health and—although they may indirectly contribute to economic development in the longer run—entail a significant and ever-increasing stream of recurring government expenditure after their completion.

An obvious way of remedying these defects is to formulate a more systematic plan of the public investment program as an integrated whole. In order to do this, it is necessary to begin by making a careful estimate of the total amount and time pattern of the financial resources that the government expects to receive during the plan period from domestic sources and from external loans and aid. Next, it is necessary to make realistic estimates of the costs and benefits of the alternative investment projects within the public sector as a whole so as to select the most productive combination of projects, taking into account significant complementary relationships between the different projects. In selecting the best combination of projects to be included in the plan, it is necessary to pay special attention to the time pattern of costs and benefits. A poor country, with limited sources of government revenue, would have to discount future benefits heavily relative to the more immediate benefits and would have to give priority to the type of project with quicker returns in the form of expansion in output and tax yields over the type of project that may promise higher rates of return, but only in the more distant future.

The problems of carrying out an integrated public investment program serve to emphasize the crucial role of the annual budget in development planning. At the aggregate level, with a given amount of external aid, the stream of the total investable funds available to the government during the plan period depends on its ability to raise revenue (and borrow from domestic sources) and, equally important, to control its nondevelopment, or "consumption," expenditure year by year during the plan period. At the individual project level, the fact that a project requires a number of budget years to complete does not dispense with the need for annual budgetary controls to ensure that it is being implemented in stages, according to the timetable as originally planned. Indeed, it is only through the discipline of annual budgetary controls that a medium-term development plan is likely to be kept nearer the course as originally planned.

Few developing countries have submitted themselves to the budgetary discipline necessary for implementing an integrated public investment program. This has not deterred them, however, from jumping from a simple type of development plan to "comprehensive" economic planning, embracing both the public and the private sectors and regulating both the aggregate level of economic activity and its detailed composition. The drive toward comprehensive planning arises from various causes: from a distrust of the automatic working of the market mechanism and its ability to promote economic development; from a desire to assert national economic independence by government control of foreign trade and investment; and from the theories of economic development, fashionable during the 1950s, that emphasize the need for a "big push" to overcome technical indivisibilities and the need for a simultaneous setting up of a number of mutually supporting projects to enjoy the benefits of technical complementaries. The economic development plans published by the developing countries in the 1960s were fairly elaborate. The trend to "quantitative" planning encouraged the use of elaborate statistical estimates and projections even when the primary statistical sources on which these computations were based were often unreliable or conjectural. Advanced mathematical techniques were also increasingly employed.

Planning the whole economy

Basically, there are three parts to such a development plan: (1) the target figures for increase in per capita income and consumption to be attained at the end of the plan (with estimated figures for the intermediate years during the plan); (2) estimates of the quantities of various resources, such as capital, manpower, and foreign exchange, needed to implement the target figures (including the time profile of the rate at which these resources will be required during the plan); and (3) parallel but independent estimates and projections of the quantities and the time pattern of these resources expected to be available both to the government and to the economy as a whole during the plan period. The elaborate planning documents issued by some developing countries may be described as attempts to quantify as far as possible the information required under the three heads and to test the formal consistency of the plan. This essentially consists in asking (a) whether the total amount of available resources is sufficient to meet the total requirements of resources as set by the target figures, and (b) whether the allocation of resources planned for different sectors is consistent with the detailed target figures for the increased output of different goods and services required for consumption and investment. When the resources required by some industries are intermediate goods (the output of other industries), input–output tables are frequently used to check whether the outputs of different industries are sufficient to supply not only the target figures for final use in the form of consumption and investment but also the "indirect use" required by other industries. The more advanced planning models using programming techniques in an attempt to solve the further question (c) whether the planned pattern of allocating resources is the most efficient; *i.e.*, whether it minimizes the resources needed to meet the target figures as compared with other patterns.

Difficulties in development planning. In spite of increasing professionalism in the formulation of development plans on paper, the practical performance of the developing countries in implementing development plans of any complexity has not been very encouraging. Development plans, however elaborately formulated on paper, rarely get beyond the first and most obvious practical hurdle, namely, how to equate the total amount of investable resources required to fulfill the target rates of economic development set by the plan with the total supply of investable resources. This arises from the practice of starting with some minimum "politically acceptable" target rate of economic growth while optimistically assuming that the problem of providing the necessary resources will somehow look after itself. This is the opposite of realistic planning, which should start from the supply of available resources and find out the maximum possible rate of economic growth that can be got out of these resources. There also appears to be a tendency to be overly optimistic as to the availability of resources and to underestimate the costs of projects. Very often, development plans have been cut short in midstream as balance of payments difficulties arising from this optimism have led the authorities to curtail their efforts sharply. Sometimes, also, the plan may be deliberately drawn up larger than can be sustained out of the available domestic resources as an elaborate window-dressing exercise to obtain a greater amount of external aid. When external aid fails to fill the planned gap in resources, the typical reaction is not to reduce the size of the plan but to take the politically less painful (and the administratively simpler) expedient of keeping to the publicly declared target rates of the plan and then trying to fill the gap in resources out of "forced saving," which it is hoped will be generated by budget deficits and inflation. Unfortunately this "forced saving" approach has not worked in most developing countries, because the

The need for realistic targets

public soon loses confidence in the stability of the purchasing power of money as prices tend to rise in step with increases in government expenditure. The pressure of domestic inflation increases the pressure of demand for imports, while rising domestic production costs discourage expansion of exports.

This disequilibrium situation may be eased by raising the rate of interest to reduce the demand for investable funds and to encourage saving, and by devaluation of the currency to discourage imports and encourage exports. Some governments have done so. But many governments in developing countries generally prefer to maintain artificially low rates of interest and to supply cheap loans to the public sector and to some favoured sections of the private sector engaged in modern manufacturing industry. Many are also reluctant to devalue for fear of further raises in prices through speculation and the higher costs of imported goods. Thus many tend to rely heavily on detailed administrative controls and import licensing to ration the scarce supply of investable funds and foreign exchange. The attempt to control the entire economy in detail through a network of direct administrative controls inevitably results in inefficiency and delays, aggravated by the inadequacy of the administrative machinery and a shortage of competent civil servants. The closer the "integration" planned between different sectors of the economy, the greater the damage to efficient coordination, since delay in one sector causes widespread delays in others.

The main weaknesses of the formal "quantitative" economic development plans are that they distract attention from a variety of important qualitative factors and focus on physical quantities rather than incentives. Qualitative factors include such matters as the practical capacity of the administrative machinery to implement the plan, the degree of political stability, and the extent of public confidence in the government's willingness and ability to carry out stated aims, which are crucial for the practical success of planning. Focus on physical quantities seems to result from a perceived need to indicate target levels of output of individual commodities; it distracts from the important fact that much economic activity is undertaken within the private sector and is responsive to incentives. Plans have tended to induce efforts to implement controls over private-sector activities. These have been effective to a large extent in preventing unplanned production activities, but they have been less effective in inducing desired increases in output without the appropriate incentives. The developing countries might do much better with a less comprehensive type of planning, making a greater use of indirect controls through the market mechanism and concentrating attention on the breaking of bottlenecks, particularly to the expansion of production in agriculture and export industries. Some developing countries have in fact succeeded in attaining rapid rates of growth just by concentrating on these vital sectors of the economy without an elaborate paraphernalia of planning.

(H.My./A.O.K.)

Economic forecasting

Economic forecasting is the prediction of any of the elements of economic activity. Such forecasts may be made in great detail or may be very general. In any case, they describe the expected future behaviour of all or part of the economy and help form the basis of planning.

The importance of theory

Formal economic forecasting is usually based on a specific theory as to how the economy works. Some theories are complicated, and their application requires an elaborate tracing of cause and effect. Others are relatively simple, ascribing most developments in the economy to one or two basic factors. Many economists, for example, believe that changes in the supply of money determine the rate of growth of general business activity. Others assign a central role to investment in new facilities—housing, industrial plants, highways, and so forth. In the United States, where consumers account for such a large share of economic activity, some economists believe that consumer decisions to invest or save provide the principal clues to the future course of the entire economy. Obviously the theory that a forecaster applies is of critical importance to the forecasting process; it dictates his line of investigation, the statistics he will regard as most important, and many of the techniques he will apply.

Judgment in forecasting

Although economic theory may determine the general outline of a forecast, judgment also often plays an important role. A forecaster may decide that the circumstances of the moment are unique and that a forecast produced by the usual statistical methods should be modified to take account of special current circumstances. This is particularly necessary when some event outside the usual run of economic activity inevitably has an economic effect. For example, forecasts of 1987 economic activity in the United States were more accurate when the analyst correctly foresaw that the exchange value of the dollar would fall sharply during the year, that consumer spending would slacken, and that interest rates would rise only moderately. None of these conclusions followed from purely economic analysis; they all required judgment as to future decisions. Similarly, an economist may decide to adjust an economic forecast that was made by traditional methods to take account of other unique conditions; he may, for example, decide that consumers will alter their spending patterns because of special circumstances such as rising prices of imports or fear of threatened shortages.

Although judgment may be based on experience and understanding, it may also be no more than unconscious bias. Forecasts based on judgment cannot be subjected to the kind of rigorous checks applied to forecasts developed by the use of more objective techniques. Consequently, the most accurate and useful forecasts are likely to be those founded on essentially economic considerations and standard statistical techniques. Though they can then be modified by the application of judgment, the resulting changes should be stated explicitly enough so that anyone wishing to use a forecast will know where, and how, it has been affected by the forecaster's own judgment, or bias.

Economic forecasting is probably as old as organized economic activity, but modern forecasting got its impetus from the Great Depression of the 1930s. The effort to understand and correct the worldwide economic disaster led to the development of a vastly greater supply of statistics and also of the techniques needed to analyze them. After World War II, many governments committed themselves to maintaining a high level of employment. Most governments of the industrialized Western countries were prepared to intervene more often and more directly in economic affairs than previously. Business organizations manifested more concern with anticipating the future. Many trade associations now provide forecasts of future trends for their members, and a number of highly successful consulting firms have been formed to provide additional forecasting help for governments and businesses.

TYPES OF FORECASTING

Forecasting the GNP and its elements. Perhaps the forecasts most familiar to the public are those of gross national product and its elements. Gross national product, or GNP, is the total value of the goods and services produced in a nation. It is, therefore, a convenient and comprehensive measure for assessing changes in general economic welfare. A forecast of the GNP also provides a useful framework for more detailed forecasts of specific industries. Almost all developed nations maintain sets of national income accounts and make forecasts as well.

The GNP can be regarded as being composed of three major components: spending by government, private investment spending, and spending by consumers. Net exports (that is, exports minus imports) are also counted in the GNP but their magnitude, which may be positive or negative, is usually small. (For the nations that depend more heavily on foreign trade, like Japan after World War II, or that incur substantial imbalances in their trade accounts, like the United States in the 1980s, net exports are of course more important.)

Forecasting government spending

Government spending is usually the easiest part of the GNP to forecast. Government expenditures can be determined with a fair degree of accuracy for well over a year in advance by studying existing budgets and appro-

priations, modified to take account of new political or economic developments. Most such adjustments are relatively minor for any forecast that runs only a year or two into the future; new government programs usually have only a small effect on expenditures in the short run. An obvious exception to this is a major change in the military situation, which can drastically alter spending plans.

It is important to note that government spending, as counted in the GNP, is not the same as total budgeted expenditures. Spending gets into the GNP only when money is paid for goods—military equipment, buildings, and so on—or services, which principally means the wages and salaries of government employees. These kinds of expenditures account for only part of the government budget; the remainder represents money transferred to bondholders, other private citizens (particularly people receiving pensions), and state and local governments. These funds affect the GNP only when they are finally spent by the recipients.

Forecasting private investment

Private investment poses far more difficult forecasting problems because it reflects many thousands of individual and corporate decisions that are not recorded publicly (as government budgets are) and that can be, and often are, changed very substantially. Private investment is the most erratic of the major categories of the GNP—the most subject to "boom and bust" cycles. A good forecast of investment spending is therefore essential to an accurate appraisal of the overall economic situation.

Capital investment by business (spending for new plants and equipment) is particularly important. The incomes generated in the process of manufacturing new equipment and building new plants play a major role in increasing consumer spending during periods of expansion. But when investment slumps, employment and incomes generally also suffer, slowing the entire economy. Business investment has thus been studied with great care, and a wide variety of methods to guide forecasters have been developed, including econometric models, surveys of business investment plans, regular reports on commitments for investment, and fundamental studies of the condition of the nation's stock of capital goods (see below *Forecasting techniques*).

Business also invests in inventory—that is, goods in the process of production and finished goods not yet sold to the final consumer. Most of the time, inventories increase roughly in line with the trend of sales. If sales fall short of expectations, however, inventories tend to become excessive. Business then moves to reduce stocks by cutting back production. Such cutbacks can aggravate economic recessions; as production is reduced because of disappointing sales, incomes are thereby reduced and sales fall further, inventories must be cut even more, and so on in a downward spiral. Consequently, correct forecasts of inventory investment are both essential to good economic forecasting and also particularly difficult. Surveys of business plans to build or reduce inventory have been helpful; econometric methods have also been applied; but inventory investment remains one of the weak links in the forecasting process.

New home construction accounts for a relatively small share of the GNP, but it is important to the forecaster because home construction is a relatively volatile industry. Homebuilding activity responds quickly to changes in the availability of mortgage money from the principal mortgage lending institutions, and thus forecasters follow closely the flow of savings to these institutions and also the level of commitments that have been made to finance future construction. Information on the number of building permits issued is also helpful, as are statistics on the volume of new construction contracts.

Forecasting consumer spending

Economists used to believe that forecasting consumer spending was fairly simple; as a rule of thumb, consumers could be counted on to spend 94–95 percent of their current income and save the rest. Thus an analyst could calculate the amount of personal income generated by government spending, private investment, and past consumer spending, adjust for tax payments, and arrive at a good estimate of consumer spending. This method still works well in determining the average rate of spending over an extended period of time. But in rich countries consumers as a group are quite free to vary their spending patterns in the short run; they may at any particular time spend more than usual because they anticipate shortages or because they believe that their incomes will rise further; or they may cut back their spending if they fear that a recession is about to develop. Such variations from normal spending patterns have their main effect on durable goods such as automobiles and household appliances; spending is far more stable for nondurables (food, clothing, and the like) and for services.

Because consumers account for such a large proportion of all economic activity, a shift of just 1 or 2 percent between spending and saving can make the difference between rapid growth or recession for the entire economy. Economists now use surveys of consumer attitudes in attempting to read the mood of the public; surveys of intentions to buy durable goods have also been helpful.

Forecasting for an industry or firm. General economic conditions set the tone for all parts of the economy. Good forecasting for an industry or firm begins, therefore, with a good analysis of the overall economy. Within this framework, the analyst must then take account of the particular factors that are most important to his own industry. In some cases, the sales of an industry may correlate fairly directly with one or more of the elements of the national income and product accounts—lumber sales with home construction, for example, or sales of nondurable consumer goods with consumer income and total consumer spending. Forecasting for industries that produce basic materials usually requires a series of projections for specific markets. A steel forecast might be based on the outlook for such major steel markets as automobiles, construction, and metal containers. The basic forecast would then be adjusted for expected shifts in exports and imports of steel and for changes in inventories of steel or steel-using products.

Forecasting is most difficult for companies that produce durable goods such as automobiles, industrial equipment, and appliances and for companies that supply the basic materials for these industries. This is because sales of such goods are subject to extreme variation. In a five-year span in the early 1970s, annual sales of automobiles in the United States increased by 22 percent in one year and declined by 22.5 percent in another. Consequently, the durable goods industries in general and automobile companies in particular have developed especially complex and sophisticated forecasting techniques. In addition to careful analysis of income trends (based on a general economic forecast), automobile companies, which are particularly sensitive to competition from imports, support a number of studies of consumer attitudes and surveys of intentions to purchase automobiles.

Forecasting for an individual firm obviously begins with a forecast for the industry or industries in which it is involved. Beyond this, the analyst must determine the degree to which the company's share of each market may vary during the forecast period. Such variations can result from the introduction of a new product, the improvement of an existing product, the opening, closing, or expansion of plants, the activities of domestic or foreign competitors, a change in sales effort, or a variety of other factors. Information required to make such assessments may come in part from the company's own investment and marketing plans. Information on the activity and sales prospects of competitors is frequently collected from the firm's own salesmen. An increasing number of companies now employ sophisticated market research techniques to determine the probable reaction of their customers to new products.

Long-term forecasting. In recent years, increasing effort has been devoted to long-range forecasting for periods extending five, 10, or more years past the normal "short-term" forecast period of one or two years. Business has come to recognize the usefulness of such forecasts in developing plans for future expansion and financing.

Labour force projections

Long-range forecasts usually are based on the assumption that activity toward the end of the period will reflect normal "full" employment. Given this assumption, the overall rate of growth depends on two principal factors:

the number of people in the labour force and the rate at which productivity (output per worker) increases. The number of people of working age is known, barring some natural disaster (and excluding immigration), far into the future; they have already been born. Forecasters usually assume that productivity will continue to grow at the typical rates of recent decades. Expected technological developments, however, may alter the projected rate of change. The combination of changes in the labour force and productivity produces an estimate of the total growth rate for the economy.

A measure of total economic activity arrived at by such methods as these serves, in effect, as a control total for making long-range forecasts of the constituent elements of the economy. If estimates for spending by consumers, government, and business add up to more than the total of goods and services that can reasonably be expected, then the projection for one or more of these elements must be reduced. If the sum of the projected parts is less than the probable total, the analyst is likely to assume a shift in economic policy that will move the economy up to full employment by the end of the forecast period and adjust his various projections up to the appropriate total.

Long-range forecasts for individual parts of the economy depend on many of the same factors as do short-range forecasts, except that cyclical factors are usually ignored. Over the longer range, however, additional factors enter. Among the most obvious of these are growth in the population and shifts in its age composition. Changes in age composition have had a major effect on both consumer and government spending patterns in many countries since World War II. The unusually large age cohorts born in the years following World War II had enormous influence on patterns of consumption and on labour-force composition. As young adults they tended to buy large amounts of durable goods and to add to the need for home construction; on the average, they saved less and borrowed more in relation to their incomes than most older people had. Their children constituted a secondary "baby boom," who could expect to see their parents become the largest generation of retired persons ever known.

In addition to population pressures, a number of other trends and assumptions influence long-range forecasts. Assumptions about war and peace are obviously critical. Assumptions must be made about government spending programs; expensive new programs may bring higher taxes and less consumer spending, whereas slower growth in government spending may lead to tax reductions. Over longer periods of time, technological discoveries or changes in financial institutions can affect the overall economy. When the forecast is made for an industry or a firm, the expected introduction of new products is also important.

FORECASTING TECHNIQUES

Economic forecasters have a vast array of information to work with and a growing variety of techniques. A few economists, believing that just one or two key factors determine the future course of the economy, limit their observations to these factors and develop forecasts based on them. A leading example of this is found in the school of thought that ascribes most importance to changes in the money supply. But most economists use a wider range of material.

Information on spending. Some elements of the future are known with reasonable accuracy. Government spending is reflected in existing budgets. These budgets indicate how much will be spent and how much money will be extracted from the stream of private spending by taxation. Similar information is available on some parts of the private economy. Periodic surveys conducted both by government and by private organizations measure business plans to invest in new plants and equipment. Increasingly, attempts are made to probe the mood and intentions of consumers concerning the possible purchase of automobiles, houses, appliances, and other durable goods. Regular surveys are also made to determine the general mood of the public—whether people are optimistic or pessimistic about their own economic future and thus whether their spending is apt to be relatively strong or relatively weak.

Surveys of business and consumer intentions

In general, such information obtained from the various surveys of investment plans, spending plans, and attitudes has been highly useful to economic forecasters. Such information helps to limit the range of possibility. But plans and attitudes change, sometimes quite abruptly, and although the surveys are useful tools they are not clear and reliable guides to the future.

Selection of turning points. Probably the single most difficult economic forecasting problem is to pick the turning points in economic activity—the times at which the economy turns from growth to recession or from recession to recovery. Because of the difficulty and importance of the problem, major efforts have been made to develop tools for this purpose. The National Bureau of Economic Research in the United States has identified a number of statistical series that normally turn up or down before the economy does. Common stock prices, business inventories, and changes in consumer installment debt are among these series, which are known as "leading indicators." Other statistical series normally move in line with overall activity ("coincident indicators"), and a third group changes direction after the economy does ("lagging indicators"). Although careful study of these groupings of statistics can be helpful, unfortunately there is no fixed relation between the movements of the economy and those of the various indicators. Although the "leading indicators" do ordinarily lead, the length of the lead varies. This reflects the dynamic nature of a complex economy that is constantly changing and in which strength or weakness may come from a variety of sources.

Some economists also use sets of statistics called diffusion indexes to calculate economic turning points. A diffusion index is a method of summarizing the common tendency of a group of statistical series. If a greater number of the series are rising than are declining, the index will be above 50; if fewer are rising than declining, it will be below 50. In effect, a diffusion index measures the degree to which either strength or weakness pervades the economy. If, for example, most of a group of industries are increasing their production rates, the economy as a whole is probably expanding; if the proportion of industries that are growing begins to decline and falls significantly below 50 percent for a period of time, the economy is probably in a recession, or at least moving in that direction.

Econometric methods

Economists frequently use mathematical equations to express the normal relations between various economic factors. As a simple example, a given increase in consumer income will ordinarily produce a certain increase in sales, saving, and tax revenue, and these developments can be expressed mathematically. With a sufficient number of equations, all the important interactions within the economy can be simulated in a mathematical model. With the advent of computers able to make millions of calculations in a few moments, economists began to construct more and more complex sets of equations, called econometric models. These models, some of which include hundreds of equations, can be used to forecast overall economic activity (macroeconomic forecasting) or developments in particular parts of the economy (microeconomic forecasting). The success of econometric forecasting has so far been limited because the exact nature of economic relations is not fully known, and also because of the inadequacies of existing statistics. Nevertheless, the improvement of these techniques represents the greatest hope for more accurate economic forecasting in the future.

Input–output analysis

The computer has also stimulated development of another potential forecasting tool, input–output analysis. Input–output tables show the relations between the various industries and sectors of an economy. They show, for each industry, the amount of its output that goes to every other industry to be used as raw materials or semifinished products, as well as the amount that goes to the final markets of the economy. Input–output tables also show each industry's consumption of the products of other industries, as well as each industry's contribution to the production process. With such a table it is possible to trace the effects of changes in one industry or sector upon all the other industries and sectors.

The usefulness of input–output analysis for forecasting

purposes has been limited by a number of factors. One problem is that it takes years to put such complex sets of statistics together; in a changing economy, relations may have shifted by the time the data for a base period have been assembled. Progress has been made, however, in developing methods to bring these relations (called technical coefficients) up to date, and input–output analysis shows increasing promise as a forecasting tool.

THE ACCURACY OF ECONOMIC FORECASTS

Major improvements have been made in the accuracy of economic forecasting. A competent economist can usually predict accurately enough to provide guidance to those who make policy decisions. Consensus forecasts—the average of a group of forecasts made by different individuals or organizations—have come closer and closer to the mark in recent years. Errors persist, nevertheless, and they occasionally lead to bad decisions.

Sources of error

The sources of error in economic forecasts are many. Some lie outside the realm of economic analysis; wars, agricultural or other natural disasters, or political upheavals are examples. Some forecasts go wrong for essentially ideological reasons: people who do not believe that an economic system will function tend to forecast its failure, which accounts for the many predictions of another great depression. Adherents of a political party in power have a notable tendency to optimism, whereas their opponents, including economists, tend to view the future with alarm. The student of forecasts must obviously consider their source; purity of motive is an important virtue for economists.

The most vexing sources of error, however, lie within the realm of economic knowledge. Many are statistical. Not only are some of the published data inaccurate but even the best statistics are available only after a period of time; the forecaster is forever predicting the future when he cannot be completely sure of the present. Statistics of inventories, among the most volatile economic elements, are noteworthy in this respect.

The most persistent form of error in economic forecasting, however, is probably theoretical. Man's knowledge of his own economic institutions is limited. Good analysis is made more difficult by the fact that these institutions are constantly changing. This means that economic theory based on experiences of the 1950s may be of limited use in the 1980s. Some of the greatest contributions to the continued improvement of economic forecasting may come from economists who are not necessarily forecasters themselves but have the insight to understand the changing economy of today.

(R.W.E.)

BIBLIOGRAPHY

Economic growth: SIMON KUZNETS, *Capital in the American Economy: Its Formation and Financing* (1961), a description of trends in capital formation in different sectors of the American economy from the mid-19th to the mid-20th century; W.W. ROSTOW, *The Stages of Economic Growth*, 2nd ed. (1971); EZRA J. MISHAN, *The Cost of Economic Growth* (1967), a critique of rapid economic growth as a policy goal; JOHN KENNETH GALBRAITH, *The Affluent Society*, 4th ed. (1984), a critique of modern capitalism written for the layman; JOSEPH A. SCHUMPETER, *Business Cycles: A Theoretical, Historical, and Statistical Analysis of the Capitalist Process*, 2 vol. (1939, reprinted 1982), a pioneering work in the field of growth and cycles that stresses the importance of entrepreneurship; and JAMES S. DUESENBERRY, *Business Cycles and Economic Growth* (1958, reprinted 1977), a theoretical study of a developed economy of the American type, with emphasis on the importance of demand in growth. Later studies include DAN USHER, *The Measurement of Economic Growth* (1980), an explanation of the difficulties inherent in such measurement; JOHN CORNWALL, *The Conditions for Economic Recovery: A Post-Keynesian Analysis* (1983), an analysis of the causes and consequences of economic stagnation; MARTIN RICKETTS, *The New Industrial Economics: An Introduction to Modern Theories of the Firm* (1987), stressing the role of entrepreneurship in modern economics; and on government intervention, STEPHEN WILKS and MAURICE WRIGHT (eds.), *Comparative Government-Industry Relations: Western Europe, the United States, and Japan* (1987).

(J.L.C.)

Economic development: The statistics of national incomes and rates of growth are supplied in the *World Development Report 1987* (1987), published for the World Bank. Conceptual problems of national income measurement are discussed in SIMON KUZNETS, *Modern Economic Growth: Rate, Structure and Spread* (1966); and GERALD M. MEIER, *Leading Issues in Economic Development*, 4th ed. (1984). For specific problems of developing countries, see H. MYINT, *The Economics of the Developing Countries*, 5th ed. (1980); IAN M.D. LITTLE, *Economic Development: Theory, Policy, and International Relations* (1982); MARTIN FRANSMAN (ed.), *Machinery and Economic Development* (1986); and GRAHAM BIRD, *International Financial Policy and Economic Development: A Disaggregated Approach* (1987). The role of foreign exchange is studied in ANNE O. KRUEGER (ed.), *Trade and Employment in Developing Countries*, 3 vol. (1981–83). For a good exposition of growth economics, see ROBERT M. SOLOW, *Growth Theory*, enl. ed. (1988); JOHN HICKS, *Capital and Growth* (1965, reissued 1972); and JOHN WORONOFF, *Asia's "Miracle" Economies* (1986).

(H.My./A.O.K.)

Economic productivity: General works dealing with productivity and its measurement include JOHN W. KENDRICK, *Understanding Productivity* (1977); JOHN W. KENDRICK and ELLIOT S. GROSSMAN, *Productivity in the United States* (1980); JEAN FOURASTIÉ, *La Productivité*, 10th ed. (1980); and GERHART E. REUSS, *Produktivitätsanalyse: Ökonomische Grundlagen und statistische Methodik* (1960). See also UNITED STATES. BUREAU OF LABOR STATISTICS, *Productivity: A Selected Annotated Bibliography* (irregular); and *Trends in Multifactor Productivity, 1948–81* (1983), updated annually by the news release *Multifactor Productivity Measures.* Estimates of the growth of output, inputs, and productivity are presented in the *OECD Economic Outlook* (semiannual), providing coverage for major countries of the world. In EDWARD F. DENISON, *Trends in American Economic Growth, 1929–1982* (1985), the author uses the "growth accounting" method that he pioneered. An early work on international comparisons of output and productivity is COLIN CLARK, *The Conditions of Economic Progress*, 3rd ed. (1957, reprinted 1983). Later works include EDWARD F. DENISON, *Why Growth Rates Differ: Postwar Experience in Nine Western Countries* (1967)—updated by JOHN W. KENDRICK, "International Comparison of Recent Productivity Trends," in WILLIAM FELLNER (ed.), *Essays in Contemporary Economic Problems* (1981); ANGUS MADDISON, "Growth and Slowdown in Advanced Capitalist Economies: Techniques of Quantitative Assessment," *Journal of Economic Literature*, 25(2):649–98 (June 1987), and the same author's *Phases of Capitalist Development* (1982). World-wide comparisons are made in IRVING A. KRAVIS and ROBERT E. LIPSEY, "The Diffusion of Economic Growth in the World Economy, 1950–80," in JOHN W. KENDRICK (ed.), *International Comparisons of Productivity and Causes of the Slowdown* (1984). The convergence thesis is examined in WILLIAM J. BAUMOL, "Productivity Growth, Convergence, and Welfare: What the Long-Run Data Show," *The American Economic Review*, 76(5):1072–85 (December 1986).

(Ma.F./J.W.K.)

Economic planning: (*Planning in Communist countries*): General works on Soviet-type planning include ALEC NOVE, *The Soviet Economy*, 3rd ed. (1986); PAUL R. GREGORY and ROBERT C. STUART, *Soviet Economic Structure and Performance*, 3rd ed. (1986); MICHAEL ELLMAN, *Socialist Planning* (1979); MARIE LAVIGNE, *The Socialist Economies of the Soviet Union and Europe* (1974; originally published in French, 1970); TREVOR BUCK and JOHN COLE, *Modern Soviet Economic Performance* (1987); DAVID A. DYKER, *The Future of the Soviet Economic Planning System* (1985); ABRAM BERGSON and HERBERT S. LEVINE (eds.), *The Soviet Economy: Toward the Year 2000* (1983); and P.J.D. WILES, *Economic Institutions Compared* (1977). For historical background, see ALEC NOVE, *An Economic History of the U.S.S.R.* (1969, reprinted with revisions, 1982); and the highly detailed and well-documented survey of the creation of the system in EDWARD HALLETT CARR and R.W. DAVIES, *Foundations of the Planned Economy, 1926–29*, 3 vol. in 5 (1969–76). For a discussion of Mikhail Gorbachev's reform program, see ABEL AGANBEGYAN, *The Challenge: Economics of Perestroika*, trans. from Russian (1988). Fundamental theoretical discussions are offered in JÁNOS KORNAI, *Economics of Shortage*, trans. from Hungarian, 2 vol. (1979, reissued 1980); V.V. NOVOZHILOV, *Problems of Cost-Benefit Analysis in Optimal Planning* (1970; originally published in Russian, 1967); JOSEPH S. BERLINER, *The Innovation Decision in Soviet Industry* (1976); and on a more empirical level, RONALD AMANN, JULIAN COOPER, and R.W. DAVIES (eds.), *The Technological Level of Soviet Industry* (1977).

On social policy, see ALASTAIR MCAULEY, *Economic Welfare in the Soviet Union* (1979). On foreign trade, see FRANKLYN D. HOLZMAN, *Foreign Trade Under Central Planning* (1974); and PHILIP HANSON, *Trade and Technology in Soviet-Western Relations* (1981). Analyses of agricultural development include

STEFAN HEDLUND, *Crisis in Soviet Agriculture* (1984); D. GALE JOHNSON and KAREN MCCONNELL BROOKS, *Prospects for Soviet Agriculture in the 1980s* (1983); and KARL-EUGEN WÄDEKIN, *Agrarian Policies in Communist Europe* (1982). More general works on socialist planning are WŁODZIMIERZ BRUS, *The Economics and Politics of Socialism* (1983); ALEC NOVE, *The Economics of Feasible Socialism* (1983); and BRANKO HORVAT, *The Political Economy of Socialism: A Marxist Social Theory* (1982). On other eastern European countries, see DAVID GRANICK, *Enterprise Guidance in Eastern Europe: Comparison of Four Socialist Economies* (1975); LJUBO SIRC, *The Yugoslav Economy Under Self-Management* (1979); and JÁNOS KORNAI, *Contradictions and Dilemmas: Studies on the Socialist Economy and Society* (1986; originally published in Hungarian, 1983), on Hungary. On the background to Chinese reforms, see MARK SELDEN and VICTOR LIPPIT (eds.), *The Transition to Socialism in China* (1982).

(A.No.)

(*Planning in developed countries*): A historical introduction to the development of economic planning in western Europe is provided in M.M. POSTAN, *An Economic History of Western Europe, 1945–1964* (1967). A major technique of planning is discussed in HAROLD A. HOVEY, *The Planning-Programming-Budgeting Approach to Government Decision-Making* (1968). Techniques of measuring social aspects of economic growth are explored in ELEANOR BERNERT SHELDON and WILBERT E. MOORE (eds.), *Indicators of Social Change: Concepts and Measurements* (1968). Later assessments of the appropriate role of government with respect to planning include SAMUEL BRITTAN, *The Role and Limits of Government: Essays in Political Economy* (1983, reissued 1987); LEO PLIATZKY, *Paying and Choosing: The Intelligent Person's Guide to the Mixed Economy* (1985); and ALICE M. RIVLIN (ed.), *Economic Choices 1984* (1984). A general survey of problems faced by governments in improving overall economic performance is given in *Structural Adjustment and Economic Performance* (1987), a collection of data prepared for the Organisation for Economic Co-operation and Development.

(Jo.Hac.)

(*Planning in developing countries*): A good introduction to the theory and practice of economic planning in the developing countries is W. ARTHUR LEWIS, *Development Planning: The Essentials of Economic Policy* (1966), which illustrates the various stages of drawing up a consistent development plan but also emphasizes that sound fundamental economic policies are more important than formal planning techniques. ALBERT WATERSTON, *Development Planning: Lessons of Experience* (1965, reissued 1974), is a well-documented survey of development plans in various countries, particularly good in discussing the various administrative and fiscal problems of development planning. WOLFGANG F. STOLPER, *Planning Without Facts: Lessons in Resource Allocation from Nigeria's Development* (1966), is a case study with a pragmatic and general approach to development planning. Later research includes WILLIAM R. CLINE and SIDNEY WEINTRAUB (eds.), *Economic Stabilization in Developing Countries* (1981), essays on a range of topics; DAVID BEVAN *et al.*, *East African Lessons on Economic Liberalization*, (1987); a study of the effectiveness of economic incentives; PAUL M. LUBECK (ed.), *The African Bourgeoisie: Capitalist Development in Nigeria, Kenya, and the Ivory Coast* (1987), a study of the current state of African economies; and NORMAN GEMMEL (ed.), *Surveys in Development Economics* (1987), a review of contemporary opinion on the subject.

(H.My./A.O.K.)

Economic forecasting: Various of the economic indicators commonly used for forecasting are surveyed in KENNETH C. LAND and STEPHEN H. SCHNEIDER (eds.), *Forecasting in the Social and Natural Sciences* (1987); JOHN LLEWELLYN, STEPHEN POTTER, and LEE SAMUELSON, *Economic Forecasting and Policy—the International Dimension* (1985); and GEOFFREY H. MOORE, *Business Cycles, Inflation, and Forecasting*, 2nd ed. (1983). Techniques for selecting appropriate data and models for use in economic forecasting are discussed in GILES KEATING, *The Production and Use of Economic Forecasts* (1985); LAWRENCE R. KLEIN and RICHARD M. YOUNG, *An Introduction to Econometric Forecasting and Forecasting Models* (1980); STEVEN C. WHEELWRIGHT and SPYROS MAKRIDAKIS, *Forecasting Methods for Management*, 4th ed. (1985); and NORMAN FRUMKIN, *Tracking America's Economy* (1987). One of the most useful sources of timely statistical data for U.S. economic forecasts is *Survey of Current Business* (monthly).

(R.W.E./Ed.)

Economic Systems

Economic systems refer to the ways in which humankind has arranged for its material provisioning. One would think that there would be a great variety of such systems, each corresponding to the many cultural arrangements that have characterized human society. Surprisingly, that is not the case. Although a wide range of institutions and social customs have been associated with the economic activities of society, only a very small number of basic modes of provisioning can be discovered beneath this variety. Indeed, history has produced but three such kinds of economic systems—those based on the principle of tradition, those organized according to command, and the rather small number, historically speaking, in which the central organizing form is the market.

The very paucity of fundamental modes of economic organization calls attention to a central aspect of the problem of economic "systems"; namely, that the objective to which all economic arrangements must be addressed has itself remained unchanged throughout human history. This unvarying objective is the coordination of the individual activities associated with provisioning—activities that range from hunting and gathering in primitive societies to administrative or financial tasks in modern industrial systems. What may be called "the economic problem" is the orchestration of these activities into a coherent social whole—coherent in the sense of providing a social order with the goods or services it requires to assure its own continuance and to fulfill its perceived historic mission.

Social coordination can in turn be analyzed as two distinct tasks. The first of these is the production of the goods and services needed by the social order, a task that requires the mobilization of society's resources, including its most valuable and most resistive resource, human effort. Of nearly equal importance is the second task, the appropriate distribution of the product. This task must not only provide for the continuance of society's labour supply (even slaves must be fed) but must also accord with the prevailing values of different social orders, all of which favour some recipients of income over others—men over women, aristocrats over commoners, property owners over nonowners, or party members over nonmembers.

All modes of accomplishing the basic tasks of production and distribution rely on social rewards or penalties of one kind or another. Tradition-based societies depend largely on communal expressions of approval or disapproval. Command systems utilize the open or veiled power of physical coercion or punishment, or the bestowal of wealth or prerogatives. The third method—the market—also brings pressures and incentives to bear, but the stimuli of gain and loss are not usually within the control of any one person or group of persons. Instead, they emerge from the "workings" of the system itself, and, on closer inspection, those workings turn out to be nothing other than the efforts of individuals to gain pecuniary rewards by supplying the things that others are willing to pay for.

There is a paradoxical aspect to the manner in which the market resolves the economic problem. In contrast to the conformity that guides traditional society or the obedience to superiors that orchestrates command society, behaviour in a market society is mostly self-directed and seems, accordingly, an unlikely means for achieving social integration. Yet, as economists ever since Adam Smith have delighted in pointing out, the clash of self-directed wills becomes converted into just such a means within the setting of competition that is an indispensable legal and social precondition for a market system to operate. The unintended outcome of this competitive engagement of self-seeking individuals is the creation of the third, and by all odds the most remarkable, of the three modes of solving the economic problem.

Not surprisingly, these three principal solutions are distinguished by the distinct attributes they impart to their respective societies. The coordinative mechanism of tradition, resting as it does on the perpetuation of social roles, is marked by a characteristic changelessness in the societies in which it is dominant. Command systems, on the contrary, are marked by their capacity to mobilize resources and labour in ways far beyond the reach of traditional societies, so that societies with command systems typically boast of large-scale achievements such as the Great Wall of China or the Egyptian pyramids.

The third of the systems, that in which the market mechanism plays the role of energizer and coordinator, is in turn marked by yet another historical attribute, resembling neither the routines of traditional systems nor the highly personalized achievements of command systems. The mark of the market is the galvanic charge imparted to economic life from the energies unleashed by its competitive, gain-oriented setting. This charge is dramatically illustrated by the trajectory of capitalism, the only social order in which the market mechanism has played a central role. In *The Communist Manifesto,* published in 1848, Karl Marx and Friedrich Engels wrote that in less than a century the capitalist system had created "more massive and more colossal productive forces than have all preceding generations together." They also wrote that it was "like the sorcerer, who is no longer able to control the powers of the nether world whom he has called up by his spells." That creative, revolutionary, and often disruptive capacity of capitalism can be traced in no small degree to the market system that performs its coordinative task.

For coverage of related topics in the *Macropædia* and *Micropædia,* see the *Propædia,* section 531, and the *Index.*

This article is divided into the following sections:

Historical development of economic systems 908
  Primitive economic systems
  Centralized states
  Preconditions for market society
The evolution of capitalism 910
  From mercantilism to commercial capitalism
  From commercial to industrial capitalism
  From industrial to state capitalism
Centrally planned systems 912
  Soviet planning
  Mixed economies
Appraising economic systems 913
  The socialist alternative
  Problems of capitalism
Bibliography 915

HISTORICAL DEVELOPMENT OF ECONOMIC SYSTEMS

Primitive economic systems. Although economics is primarily concerned with the modus operandi of the market mechanism, an overview of premarket coordinative arrangements is not only of interest in itself but also throws a useful light on the distinctive properties of market-run societies. The earliest and by far the most historically numerous of economic systems has been that of primitive society with its central order-bestowing agency of tradition. Such economic forms of social organization must be far more ancient than Cro-Magnon man, and a few are still preserved in Eskimo tribes or bands of Kalahari hunters or Bedouin tribespeople. So far as is known, all tradition-bound peoples solve their economic problems today much

as they did 10,000 years or perhaps 10,000 centuries ago—adapting by migration or movement to changes in season or climate, sustaining themselves by hunting and gathering or by slash-and-burn agriculture, and distributing their output by reference to well-defined social claims. Elizabeth Marshall Thomas describes this distributive system in *The Harmless People:*

> It seems very unequal when you watch Bushmen divide the kill, yet it is their system, and in the end no person eats more than the other. That day Ukwane gave Gai still another piece because Gai was his relation. Gai gave meat to Dasina because she was his wife's mother. . . . No one, of course, contested Gai's large share, because he had been the hunter. . . . No one doubted that he would share his large amount with others, and they were not wrong, of course; he did.

Subsistence in primitive economies

Besides the inertial property that is perhaps the outstanding attribute of these primitive economic societies, two further aspects deserve attention. The first concerns their level of subsistence, long deemed to have been one of chronic scarcity and want. According to the still controversial findings of the anthropologist Marshall Sahlins this is not true. His studies of several primitive peoples found that they could easily increase their provisioning if they so desired. The condition usually perceived by nonprimitive observers as scarcity is felt by them as satiety: Sahlins describes primitive life as the first "affluent society."

A second attribute of interest in primitive economic systems is the difficulty of describing any part of their life activities as constituting an "economy." No special modes of coordination set the activities of hunting or gathering, or the procedures of distribution, apart from the rest of social life, so that there is nothing in Kalahari or Eskimo or Bedouin life for which there is needed a special vocabulary or conceptual apparatus called "economics." The economy as a network of provisioning activities is completely absorbed within and inextricable from the traditional mode of existence as a whole.

Centralized states. Very little is known of the origin of the second of the great systems of social coordination, namely the creation of a central apparatus of command and rulership. From dimly perceived clusters of population, impressive civilizations emerged in Egypt, China, and India during the 3rd millennium BC, bringing with them not only dazzling advances in culture but also the enormously powerful instrument of state power as a new moving force in history.

The appearance of these centralized states is arguably the single most decisive alteration in economic, and perhaps in all, history. Although tradition still exerted its stabilizing and preserving role at the base of these societies—Adam Smith said that in "Indostan or ancient Egypt . . . every man was bound by a principle of religion to follow the occupation of his father"—the vast temple complexes, irrigation systems, fortifications, and cities of ancient India and China, or of the kingdoms of the Inca and Maya, attest unmistakably to the difference that the organizing principle of command brought to economic life. It lay in the ability of centralized authority to wrest considerable portions of the population away from their traditional occupations and to use their labour energies in ways that expressed the wishes of a ruling personage or small elite.

Herodotus recounts how the pharaoh Khufu used his power to this end:

> [He] ordered all Egyptians to work for himself. Some, accordingly, were appointed to draw stones from the quarries in the Arabian mountains down to the Nile, others he ordered to receive the stones when transported in vessels across the river. . . . And they worked to the number of a hundred thousand men at a time, each party during three months. The time during which the people were thus harassed by toil lasted ten years on the road which they constructed, and along which they drew the stones; a work, in my opinion, not much less than the Pyramids.

Creation of economic surplus

The creation of these monuments illustrates an important general characteristic of all systems of command. It is that such systems, unlike those based on tradition, can generate immense surpluses of wealth—indeed, the very purpose of a command organization of economic life can be said to lie in the generation of such a surplus. Command systems thereby acquire the wherewithal to change the conditions of material existence in far-reaching ways. Until the modern era, when command has become the main coordination system for socialism, it was typical of such systems to use this productive power principally to cater to the consumption or to the power and glory of their ruling elites.

Moral judgments aside, this highly personal disposition of surplus has the further consequence of again resisting any sharp analytical distinction between the workings of the economy of such a society and that of its larger social framework. The methods of coordination that create and distribute wealth in a command system are identical with those that guide the imperial state in all its historical engagements, just as in primitive society the methods that coordinate the activities of production and distribution are indistinguishable from those that shape family or religious or cultural life. Thus, in command systems, as in tradition-based ones, there is little or no need for a special kind of social analysis called economics.

Preconditions for market society. These general considerations throw into relief the nature of the economic problems that must be resolved in a system of market coordination. Such a system must be distinguished from the mere existence of marketplaces. Marketplaces have existed far back in history. Trading relations between the ancient Levantine kingdoms and the pharaohs of Egypt in about 1400 BC are known from the tablets of Tell el-Amarna. A thousand years later Isocrates boasted of the thriving trade of classical Greece, and a rich and varied network of commodity exchange, as well as a well-established market for money capital, were prominent features of classical Rome.

These flourishing institutions of commerce testify to the ancient lineages of money, profit-mindedness, and mercantile groups, but they do not testify to the presence of a market system. In premarket societies, markets were the means to join suppliers and demanders of luxuries and superfluities, but they were not the means by which the provision of essential goods and services was assured. For these purposes ancient kingdoms or republics still looked to tradition and command, utilizing slavery as a basic source of labour (including captives taken in war) and viewing with disdain the profit orientation of market life. This disdain applied particularly to the use of the incentives and penalties of the market as a means of marshaling labour. Aristotle expressed the common feeling of his age when he declared, "The condition of the free man is that he does not live for the benefit of another." With the exception of some military service, nonslave labour was simply not for sale.

Markets and a market system

The difference between a society with flourishing markets and a market-coordinated society is not, therefore, merely one of attitudes. Before a system orchestrated by the market can replace one built on obedience to communal or superior pressure, the social orders dependent on tradition and command must be replaced by a new order in which individuals are expected to fend for themselves and in which all are permitted, even encouraged, to improve their material condition. Individuals cannot have such aims, much less such "rights," until the unchallengeable authority of custom or hierarchical privilege has been swept away. A rearrangement of this magnitude entails wrenching dislocations of power and prerogative. A market society is not, consequently, merely a society coordinated by markets. It is, of necessity, a social order with a distinctive structure of laws and privileges.

It follows that a market society requires an organizing principle that, by definition, can no longer be the respect accorded to tradition or the obedience owed to a political elite. This principle becomes the generalized search for material gain. Such a condition of universal upward striving is unimaginable in a traditional society and could only be seen as a dangerous threat in a society built on established hierarchies of authority. But, for reasons that will be seen, it is magically accommodated by, and indeed constitutive of, the workings of a market system.

The process by which these institutional and attitudinal changes are brought about constitutes a grand theme—per-

haps the grand theme—of economic history from roughly the 5th to the 18th and even into the 19th century in Europe. In terms of political history the period was marked by the collapse of the Roman empire, the rise of feudalism, and the slow formation of national states. In social terms it featured the end of an order characterized by an imperial retinue at the top and massive slavery at the bottom, its replacement by gradations of feudal vassalage descending from lord to serf, and the eventual appearance of a bourgeois society with distinct classes of workers, landlords, and capitalists. From the economist's perspective it was marked by the breakdown of a coordinative mechanism of centralized command, the rise of the mixed pressures of tradition and local command characteristic of the feudal manor, and the gradual displacement of those pressures by the material penalties and rewards of an all-embracing market network. In this vast transformation, the rise of the market mechanism became crucial as the means by which the new social formation of capitalism assured its self-provisioning, but the mechanism itself rested on deeper-lying social, cultural, and political changes that created the capitalist order which it served.

To attempt to trace these lineages of capitalism would take one far beyond the confines of the present subject. Suffice it to remark that the emergence of the new order was first given expression in the 10th and 11th centuries when a rising mercantile "estate" began to bargain successfully for recognition and protection with the local lords and impecunious monarchs of the early Middle Ages. Not until the 16th and 17th centuries was there a "commercialization" of the aristocratic strata, many of whom fared poorly in an ever more money-oriented world, and who accordingly contracted marriages with merchant families whom they would not have received at home a generation or two earlier. Of greatest significance, however, was the transformation of the lower orders, a process that began in Elizabethan England but did not take place en masse until the 18th and even 19th centuries. As traditional lords became profit-minded landlords, peasants were forced off the land to become an agricultural proletariat in search of the best wages it could get because its traditional subsistence was no longer available. Thus the market network extended its disciplinary power over "free" labour—the resource that had always previously eluded its influence. A social order had been created in which markets could coordinate production and distribution in a manner never before possible.

THE EVOLUTION OF CAPITALISM

Emergence of mercantilism

From mercantilism to commercial capitalism. It is usual to describe the earliest stages of capitalism as mercantilism, the word denoting the central importance of the merchant overseas traders who rose to prominence in 17th- and 18th-century England, Germany, and the Low Countries. In numerous pamphlets these merchants defended the principle that their trading activities buttressed the interest of the sovereign power, even when, to the consternation of the court, this required sending "treasure" (bullion) abroad. As the pamphleteers explained, treasure used in this way became itself a commodity in foreign trade, in which, as the great merchant Thomas Mun wrote about 1630, "we must ever observe this rule; to sell more to strangers than we consume of theirs in value."

For all its trading mentality, mercantilism was only partially a market-coordinated system. Smith complained bitterly about the government monopolies that granted exclusive trading rights to groups such as the East India or the Turkey companies, and modern commentators have emphasized the degree to which mercantilist economies relied on regulated, not free, prices and wages. The economic society that Smith described in *The Wealth of Nations* in 1776 is much closer to modern society, although it differs in many respects, as shall be seen. This 18th-century stage is called "commercial capitalism," although it should be noted that the word "capitalism" itself does not actually appear in the pages of Smith's great book.

Smith's society is nonetheless recognizable as capitalist precisely because of the prominence of those elements that had been absent in its mercantilist form. For example, with few exceptions, the production and distribution of all goods and services were entrusted to market forces rather than to the rules and regulations that had abounded a century earlier. The level of wages was likewise mainly determined by the interplay of the supply of, and the demand for, labour, not by the rulings of local magistrates. Profits were exposed to competition rather than protected by government monopoly.

Perhaps of greater importance in perceiving Smith's world as capitalist, as well as market-oriented, is its clear division of society into an economic and a political realm. The role of government had been gradually narrowed until Smith could describe its duties as consisting of only three functions: (1) the provision of national defense, (2) the protection of each member of society from the injustice or oppression of any other, and (3) the erection and maintenance of those public works and public institutions (including education) that would not repay the expense of any private enterpriser, although they might "do much more than repay it" to society as a whole. And if the realm of government had been greatly delimited, that of commerce had been greatly expanded. The accumulation of capital had become clearly recognized as the driving engine of the system. The expansion of "capitals"—Smith's term for firms—was the motive power by which the market system was launched on its historic course.

Narrowed role of the state

Thus in *The Wealth of Nations* it was possible for the first time to describe quite precisely the dynamics as well as the coordinative processes of capitalism. The latter were entrusted to the market mechanism—which is to say, to the universal drive for material betterment, curbed and contained by the necessary condition of competition. Smith's great perception was that the combination of this drive and counterforce would direct productive activity toward those goods and services for which the public had the means and desire to pay, while forcing producers to satisfy those wants at prices that yielded no more than normal profits. Later economists would devote a great deal of attention to the question of whether competition in fact adequately constrains the workings of the acquisitive drive and whether a market system might not display cycles and crises unmentioned in *The Wealth of Nations.* These were questions unknown to Smith, because the institutions that would produce them, above all the development of large-scale industry, lay in the future. Given these historic realities, one can only admire Smith's perception of the market as a means of solving the economic problem.

Smith also saw that the competitive search for capital accumulation would impart a distinctive tendency to a society that harnessed its motive force. He pointed out that the most obvious way for a manufacturer to gain wealth was to expand his enterprise by hiring additional workers. As firms expanded their individual operations, manufacturers found that they could subdivide complex tasks into simpler ones and could then speed along these simpler tasks by providing their operatives with machinery. Thus the expansion of firms made possible an ever finer division of labour, and the finer division of labour, in turn, improved profits by lowering the costs of production, thereby encouraging the further enlargement of the firms. In this way the incentives of the market system gave rise to the augmentation of the wealth of the nation itself, endowing market society with its all-important historical momentum and at the same time making room for the upward striving of its members.

Capitalism and productivity

One final attribute of the emerging system must be noted. This is the tearing apart of the formerly seamless tapestry of social coordination. Under capitalism two realms of authority existed where there had formerly been only one—a realm of political governance for such purposes as war or law and order and a realm of economic governance over the processes of production and distribution. Each realm was largely shielded from the reach of the other. The capitalists who dominated the market system were not automatically entitled to governing power, and the members of government were not entrusted with decisions as to what goods should be produced or how social rewards should be distributed. This new dual structure brought with it two consequences of immense importance.

The first was a limitation of political power that proved of very great importance in establishing democratic forms of government. The second, closer to the present theme, was the need for a new kind of analysis intended to clarify the workings of this new semi-independent realm within the larger social order. Thus did the appearance of capitalism give rise to the discipline now called economics.

From commercial to industrial capitalism. Commercial capitalism proved to be only transitional. The succeeding form would be distinguished by the pervasive mechanization and industrialization of its productive processes, a change that would not only vastly alter the social and physical landscape but would also set into motion new dynamic tendencies for the system.

The transformative agency was already present in Smith's day, observable in a few coal mines where steam-driven engines invented by Thomas Newcomen pumped water out of the pits. The diffusion and penetration of such machinery-driven processes of production during the first quarter of the next century has been traditionally called "the" Industrial Revolution, although historians today stress the long germination of the revolution and the many phases through which it passed. There is no doubt, however, that a remarkable confluence of advances in agriculture, cotton spinning and weaving, iron manufacture, machine tool design, and the harnessing of mechanical power began profoundly to alter the character of capitalism in the last years of the 18th century and the first decades of the 19th.

The alterations did not affect the driving motive of the system or its reliance on market forces as its coordinative principles. Their effect was rather on the social complexion of the society that contained these new technologies and on the economic outcome of the processes of competition and capital accumulation. This aspect of industrialization was most immediately apparent in the advent of the factory as the archetypal locus of production. In Smith's time the individual enterprise was still small—the opening pages of *The Wealth of Nations* describe the effects of the division of labour in a 10-man pin factory—but by the early 19th century the increasing mechanization of labour, coupled with the application of water and steam power, had raised the size of the work force in an ordinary textile mill to several hundreds, by midcentury in the steel mills to several thousands, and by the end of the century in the railways to tens of thousands.

Labour under capitalism

The increase in the scale of employment brought a marked change in the character of employment itself. In Smith's day the social distance between employer and labourer was still sufficiently small that the very word "manufacturer" implied an occupation (a mechanic) as well as a capitalist-like position. By the time of the "dark satanic mills" of the 1830s, a great gulf had opened between the manufacturers, who were now a propertied business class, and the men, women, and children who tended their machines for 10- and 12-hour stints. It was from the spectacle of mill labour, described in unsparing detail by the inspectors authorized by the first Factory Act of 1802, that Marx drew much of the indignation that animated his analysis of capitalism. More important, it was from this same factory setting, and from the urban squalor that industrialization also brought, that capitalism derived much of the social consciousness—sometimes revolutionary, sometimes reformist—that was to play so large a part in its subsequent political life.

The degradation of the physical and social landscape was the aspect of industrialization that first attracted attention, but it was its slower-acting impact on economic growth that was ultimately to be judged its most significant effect. A single statistic may dramatize this process. Between 1788 and 1839 the output of pig iron in Britain rose from 68,000 to 1,347,000 tons. For the significance of these numbers to be grasped, they must be translated into the huge multiplication of iron pumps, iron machine tools, iron pipes, iron rails, and iron beams that this increase in pig iron production made possible, and this multiplication of iron implements must in turn be translated into the strength and speed that they lent to production itself. This was the means by which the first industrial revolution promoted economic growth, not immediately but with gathering momentum. Thirty years later this effect would be repeated with even more spectacular results when the Bessemer converter ushered in the age of steel rails, ships, machines, girders, wires, pipes, and containers.

The most important consequence of the industrialization of capitalism was therefore its powerful effect on enhancing what Marx called "the forces of production"—the source of what is now called standard of living. The Swiss economic demographer Paul Bairoch has calculated that gross national product per capita in the developed countries rose from $180 in the 1750s (in dollars of 1960 purchasing power) to $780 in the 1930s and then to $3,000 in the 1980s, whereas the per capita income of the less developed countries remained unchanged at about $180–$190 from 1750 to 1930 and thereafter rose only to $410 in 1980. The single most important explanation for this difference is that the first group of countries became industrialized and the second did not.

Industrialization and national wealth

A second major effect of the industrialization of the system was its increasing economic instability. Market systems are intrinsically susceptible to perturbations, because of mismatches and miscalculations in individual markets. Such disturbances bring relatively small effects in an economy of small enterprises, but they can have major repercussions as the scale of enterprises becomes larger. The difference can be likened to the contrast between a sand pile and a girdered structure: the sand pile absorbs blows without collapsing because of the adjustment of its many small particles, while a girdered structure can fail if a single beam buckles.

Not surprisingly, then, one side effect of industrialization was the effort to minimize or prevent such shocks by linking firms together into cartels or trusts or simply into giant integrated enterprises. Although these efforts dampened the repercussions of individual miscalculations, they were insufficient to guard against the effects of speculative panics or commercial convulsions. By the end of the 19th century, economic depressions had become a worrisome and recurrent problem, and the Great Depression of the 1930s rocked the entire capitalist world. During that debacle, the gross national product (GNP) in the United States fell by almost 50 percent from its peak to its trough; business investment fell by 94 percent, while unemployment rose from 3.2 percent of the civilian labour force to 24.9 percent. Of all the reasons that account for the extraordinary increase in instability from 1830 to 1930, none is more persuasive than the increase in scale from pin factories to giant enterprises.

From industrial to state capitalism. The problem of instability takes on further importance insofar as it is a principal cause of the next structural phase of the system. The new phase is often described as state capitalism because its outstanding feature is the enlargement in size and functions of the public realm. In 1929, for example, total U.S. government expenditures—federal, state, and local—came to less than 10 percent of the GNP; from the 1970s, they amounted to roughly one-third. This increase is observable in all major capitalist nations, many of which have reached considerably higher ratios of government disbursements to GNP than the United States.

At the same time, the function of government changed as decisively as its size. Already by the last quarter of the 19th century the emergence of great industrial trusts had provoked legislation in the United States (although not in Europe) to curb the monopolistic tendencies of industrialization. Apart from these antitrust laws and the regulation of a few industries of special public concern, however, the functions of the federal government were not significantly broadened from Smith's vision. Prior to the Great Depression, for example, the great bulk of federal outlays went for defense and international relations, for general administrative expense and interest on the debt, and for the post office.

The Great Depression radically altered this limited view of government in the United States, as it had earlier begun to widen it in Europe. The provision of old-age pensions, relief for the hungry and poor, and a dole for the unemployed were all policies inaugurated by the administration

Growth of government's role

of President Franklin D. Roosevelt, following in the footsteps of similar enlargements of government functions in Britain, France, and Germany. From the 1970s forward, such new kinds of federal activities—under the designation of social security, health, education, and welfare expenditures—grew to be 20 to 50 percent larger than the traditional categories of federal spending.

Thus, one very important element in the advent of a new stage of capitalism was the emergence of the public sector as a guarantor of public economic well-being, a function that would never have entered Smith's imagination. A second and equally important departure was the government's assumption of responsibility for the general course of economic conditions themselves. This was a change of policy orientation that also emerged from the challenge of the Great Depression. Once regarded as a matter beyond remedy, the general level of national income came to be seen by the end of the 1930s as the responsibility of government, although the measures taken to improve conditions were on the whole timid, often wrongheaded (such as highly protectionist trade policies), and only modestly successful. Nonetheless, the appearance in that decade of a new economic accountability for government constitutes in itself sufficient reason to describe capitalism today in terms that distinguish it from its industrial, but largely unguided, past.

There is little doubt that capitalism in the 21st century will undergo still further structural alterations. Technological advances are rapidly reducing to near insignificance the once formidable barriers and opportunities of economic geography. Among the startling consequences of this technological leveling of the world have been the displacement of the economic centre of the globe from the Atlantic to the Pacific and the threat of large displacements of high-tech manufacture from Europe and North America to low-wage regions of Southwest Asia, Latin America, and possibly Africa. Another change has been the unprecedented growth of international finance to the point at which the total value of transactions in foreign exchange is estimated to be at least 20 times that of all foreign movements of goods and services. This boundary-blind internationalization of finance, combined with the boundary-defying ability of large corporations to locate their operations in low-wage countries, poses a challenge to the traditional economic sovereignty of nations, a challenge arising from the new capabilities of capital itself.

Internationalization of finance

A third change again involves the international economy, this time through the creation of new institutions for the management of international economic trade. A number of capitalist nations have met the challenges of the fast-growing international economy by joining the energies of the private sector (including organized labour) with the financial and negotiating powers of the state. This "corporatist" approach, most clearly evident in the organization of the Japanese economy, may become a common structural characteristic of capitalism in the 21st century, not only in managing foreign trade and production but also in dealing with domestic problems such as inflation.

These still-emergent trends seem likely to exert their pressures toward further growth of the governmental guiding role within capitalism. It is not necessary, however, to venture risky predictions as to the future. Rather, it seems more useful to conclude with two generalizations. The first emphasizes that capitalism in all its variations continues to be distinguished from other economic systems by the priority accorded to the drive for wealth and the centrality of the competitive mechanism that channels this drive toward those ends that the market rewards. The second generalization is that this driving force and constraining mechanism appear to be compatible with a wide variety of institutional settings, including substantial variations in the relationships between the private and public sectors. It is to this very adaptability that capitalism appears to owe its continued vitality.

CENTRALLY PLANNED SYSTEMS

No survey of comparative economic systems would be complete without an account of centrally planned systems, the modern descendants of the command economies of the imperial past. In sharpest possible contrast to those earlier tributary arrangements, however, modern command societies have virtually all been organized in the name of socialism—that is, with the function of command officially administered on behalf of the broad masses of the population.

Socialist central planning needs to be differentiated from the idea of socialism itself. The latter draws on moral precepts of concern for the needy that can be discovered in the Judeo-Christian tradition and derives its general social orientation from Gerrard Winstanley's Diggers movement during the English Civil Wars in the mid-17th century: "The Earth," Winstanley wrote, "was made by Almighty God to be a Common Treasury of livelihood to the whole of mankind . . . without respect of persons."

Socialism as a means of orchestrating a modern industrial system did not receive explicit attention until the Russian Revolution in 1917. In his brochure *The State and Revolution,* written before he came to power, Lenin envisaged the task of coordinating a socialist economy as little more than delivering production to central collecting points from which it would be distributed according to need—an operation requiring no more than "the extraordinarily simple operations of watching, recording, and issuing receipts, within the reach of anybody who can read and who knows the first four rules of arithmetic." After the revolution, it soon became apparent that the problem was a great deal more difficult than that. The mobilization of manpower required the complex determination of appropriate amounts and levels of pay, and the mobilization of foodstuffs from the countryside posed the awkward question of the degree to which the "bourgeois" peasantry would have to be accommodated. As civil war raged in the country, these problems intensified until production fell to a catastrophic 14 percent of prewar levels. By the end of 1920 the system was at the verge of collapse.

Lenin's simplistic view

To forestall disaster, Lenin instituted the New Economic Policy (NEP), which amounted to a partial restoration of capitalism, especially in retail trade, small-scale production, and agriculture. Only the "commanding heights" of the economy remained in government hands. The NEP resuscitated the economy but opened a period of intense debate as to the use of market incentives versus moral suasion or more coercive techniques. The debate, which remained unresolved during Lenin's life, was brought to a conclusion at his death in 1924, when Joseph Stalin rose to power and rapidly forced the collectivization of the economy. Private agriculture was converted into collective farming with great cruelty and loss of life, all capitalist markets and private enterprises were quickly and ruthlessly eliminated, and the direction of economic life was consigned to a bureaucracy of ministries and planning agencies. By the 1930s a structure of centralized planning had been put into place that was to coordinate the Russian economy for the next half century.

Soviet planning. At the centre of the planning system was the Gosplan (*gos* means "committee"), the top economic planning agency of the Soviet state. Above the Gosplan were the political arms of the Soviet government, while below it were smaller planning agencies for the various Soviet republics. The Gosplan itself was staffed by economists and statisticians charged with drawing up what amounted to a blueprint for national economic activity, usually for a five- to seven-year period. This blueprint translated the major objectives determined by political decision (electrification targets, agricultural goals, transportation networks, and the like) into industry-specific requirements (outputs of generators, fertilizers, steel rails). These general requirements were then referred to ministries charged with the management of the industries in question, where the targets were further broken down into specific outputs (quantities, qualities, shapes, and sizes of steel plates, girders, rods, wires, and so forth) and where lower-level goals were fixed, such as budgets for firms, wage rates for different skill levels, or managerial bonuses.

Planning was not, therefore, entirely a one-way process. General objectives were indeed transmitted from the top down, but as each ministry and factory inspected its obligations, specific obstacles and difficulties were transmitted

from the bottom up. The final plan was thus a compromise between the political objectives of the Central Committee of the Communist Party and the nuts-and-bolts considerations of the echelons charged with its execution. This coordinative mechanism worked quite well when the larger objectives of the system called for a kind of crash planning that resembled a war economy. The Soviet economy achieved unprecedentedly rapid progress in its industrialization drive before World War II and in repairing the devastation that followed the war. Moreover, in areas where the political stakes were high, such as space technology, the planning system was able to concentrate skills and resources regardless of cost. Yet, charged with the orchestration of a civilian economy in normal peacetime conditions, the system of centralized planning failed seriously, for reasons that will be examined below.

Perestroika in Soviet planning

Because of its failures, a far-reaching reorganization of the system was set into motion in 1985 by Mikhail Gorbachev, under the banner of *perestroika* ("restructuring"). The extent of the restructuring can be judged by these proposed changes in the coordinative system: (1) the scope and penetration of planning were to be greatly curtailed and directed mainly at general economic goals, such as rates of growth, consumption or investment targets, or regional development; (2) planning done for factory enterprises was to give way to planning by factories themselves, guided by considerations of profit and loss; (3) factory managers were no longer to be bound by instructions with respect to the sources of their inputs and the destination of their outputs but were to be free to buy from and to sell to whomever they pleased; (4) managers were also to be free to hire and—more important—to fire workers, who formerly could not be easily discharged; and (5) many kinds of small private enterprises were to be encouraged, especially in farming and the retail trades.

This program obviously represented a dramatic retreat from the original idea of central planning. One cannot say, however, that it also represented a decisive turn from socialism to capitalism, for it was not clear to what extent the restructured planning system might embody other essential features of capitalism, such as private ownership of the means of production and the exclusion of political power from the normal operations of economic life. Nor was it known to what extent economic *perestroika* was to be accompanied by its political counterpart, *glasnost* ("openness"). Thus the degree of change in both the economic structure and the underlying political order remained indeterminate. The record of *perestroika* over the rest of the 1980s was disappointing. After an initial flush of enthusiasm, the task of abandoning the centralized planning system proved to be far more difficult than anticipated, to no small degree because the achievement of such a change would have necessitated the creation of a new structure of economic (managerial) power, independent of, and to some extent in continuous tension with, that of political power, much as under capitalism. Also, the operation of the centralized planning system, freed from some of the coercive pressures of the past but not yet infused with the energies of the market, rapidly deteriorated. By the end of the decade, the Soviet system was facing an economic breakdown more severe and far-reaching than the worst capitalist crisis of the 1930s. Not surprisingly, the unrest aroused ancient nationalist rivalries and ambitions.

As the Soviet central government gradually lost control over the economy at the republic and local levels, the system of central planning eroded without adequate free-market mechanisms to replace it. By 1990 the Soviet economy had slid into near-paralysis, and this condition foreshadowed the fall from power of the Soviet Communist Party and the breakup of the Soviet Union itself into a group of independent republics in 1991. Attempts to transform socialist systems into market economies have occurred in eastern and central Europe since 1989 and in the former Soviet Union since 1992. Ambitious privatization programs have been pursued in Poland, Hungary, Germany, the Czech Republic, and Russia. The same countries also have made the transition (with varying degrees of success) to democratic forms of governance.

Mixed economies. The socialist turn from planning toward the market provides a fitting initial conclusion to this overview of the typology of economic systems, for it is apparent that the three ideal types of tradition, command, and market have never been attained in wholly pure form. Perhaps the most undiluted of these modes in practice has been that of tradition, the great means of orchestration in pre-state economic life. But, even in tradition, a form of command can be seen in the expected obedience of community members to the sanctions of tradition. In the great command systems of the past, as has been seen, tradition supplied important stabilizing functions, and traces of market exchange served to connect these systems to others beyond their borders. The market system, as well, has never existed in wholly pure form. Market societies have always taken for granted that tradition would provide the foundation of trustworthiness and honesty without which a market-knit society would require an impossible degree of supervision, and no capitalist society has ever existed without a core of public economic undertakings, of which Smith's triad—defense, law and order, and nonprofitable public works—constitute the irreducible minimum.

Experiments in mixed economies

Thus it is not surprising that the Soviet Union's efforts to find a more flexible amalgam of planning and market were anticipated by several decades of cautious experiment in some of the socialist countries of eastern Europe, especially Yugoslavia and Hungary, and by bold departures from central planning in China after 1979. All these economies existed in some degree of flux as their governments sought configurations best suited to their institutional legacies, political ideologies, and cultural traditions. All of them also encountered problems similar in kind, although not in degree, to those of the Soviet Union as they sought to escape the confines of highly centralized economic control.

Something of this mixed system of coordination can also be seen in the less developed regions of the world. The panorama of these economies is a kind of museum of economic systems, with tradition-dominated tribal societies, absolute monarchies, and semifeudal societies side by side with military socialisms and sophisticated, but unevenly developed, capitalisms. To some extent this spectrum reflects the legacy of 19th-century imperialist capitalism, against whose cultural, as well as economic, hegemony all latecomers have had to struggle. Little can be ventured as to the outcome of this astonishing variety of economic structures. A few countries may follow the corporatist model of the Pacific Rim economies; others may emulate the social democratic states of western Europe; a few will pursue a *laissez faire* approach; yet others will seek whatever method—either market or planning—that might help them establish a viable place in the international arena. Unfortunately, many are likely to remain destitute for some time. In this fateful drama, considerations of culture and politics are likely to play a more determinative role than any choice of economic instrumentalities.

APPRAISING ECONOMIC SYSTEMS

The study of economic systems begins from a historical account of the three coordinative modes of production and distribution, but it leads toward normative appraisals or comparisons of the advantages and deficiencies of the social orders that principally depend on these modes. To survey, much less to probe, these problems for capitalism and socialism would require nothing less than a review of contemporary economics. The remainder of this article will therefore briefly discuss a few aspects of socialism and of capitalism that highlight some of the major normative issues of this historical overview.

The socialist alternative. The chief economic problem of socialism has been the efficient performance of the very task for which its planning apparatus exists, namely the effective coordination of production and distribution. Conservative critics have declared that a planned economy is impossible—*i.e.*, will inevitably become unmanageably chaotic—by virtue of the need for a planning agency to make the millions of dovetailing decisions necessary to bring into existence the gigantic catalog of goods and services of a modern society. Precisely such problems became manifest in the late 1980s in the Soviet Union.

The proposed remedy for this, as noted above, is the use

of market arrangements within socialism, under which managers are free to conduct the affairs of their enterprises according to the dictates of supply and demand rather than those of a central authority. The difficulty with this solution lies in its political rather than economic requirements. The acceptance of a market system entrusted with the coordination of the bulk of economic activity requires the tolerance of a sphere of private authority apart from that of public authority. A market mechanism may be compatible with a society of socialist principles, but it requires that the present forms of such a society be radically reorganized. The political difficulties of such a reorganization are twofold. One is the tensions that can be expected to exist between the private interests, and no doubt the public visions, of the managerial echelons and those of the political regime. The creation of a market is tantamount to the creation of a realm within the socialist order into which the political arm of government is not allowed to reach fully. This almost surely entails the creation and defense of various privileges, including contract and property, that resemble those of capitalism. The threat is thus the introduction of a capitalist virus within the socialist order, the mirror image of the socialist virus that is often perceived in capitalist societies as residing in their government realms.

A second political difficulty concerns the working class, not the managerial elites. The establishment of a market system as a major coordinator of economic activity necessarily introduces the use of unemployment as a disciplining force into a social order that uses commands to allocate employment and that does not permit workers to be hired or fired for strictly economic reasons. This use of unemployment is likely to exacerbate class tensions between workers and management. Socialist reformers envisage the overcoming of these tensions by increasing worker participation in the management of the enterprises in which they work. This has been attempted for some years on a limited scale in a few countries, but no great successes have been reported.

Problems of capitalism. Advocates and critics of capitalism agree that its distinctive contribution to history has been the encouragement of economic growth. Critics, however, cite three dysfunctions that reflect its market origins.

The unreliability of growth. The first of these problems is already familiar from the above survey of the stages of capitalist development. It is the instability that has characterized and plagued the system since the advent of industrialization. Because capitalist growth is driven by profit expectations, it fluctuates with the adventitious opening of technological or social opportunities for capital accumulation. As opportunities appear, capital rushes in to take advantage of them, bringing as a consequence the familiar attributes of a boom. Sooner or later, however, the rush subsides, as the demand for the new products or services becomes saturated, bringing a halt to investment, a shakeout in the main industries caught up in the previous boom, and the advent of recession. Hence economic growth comes at the price of a succession of market gluts as booms meet their inevitable end.

Capitalism's boom-and-bust cycle

This criticism did not receive its full exposition until the publication of Marx's *Das Kapital* in 1867. For Marx, the path of growth is not only unstable for the reasons just mentioned—Marx called such uncoordinated movements the "anarchy" of the market—but increasingly unstable. The reason for this is also familiar. It is the result of the industrialization process, which leads toward large-scale enterprises. As each saturation brings growth to a halt, a process of winnowing takes place in which the more successful firms are able to acquire the assets of the less successful. Thus the very dynamics of growth tend to concentrate capital into ever-larger firms. This leads to still more massive disruptions when the next boom ends, a process that terminates, according to Marx, only when the temper of the working class snaps and capitalism is replaced by socialism.

Marx's apocalyptic expectations have been largely replaced since the 1930s by the less violent but equally disquieting views of the English economist John Maynard Keynes, first set forth in his influential *The General Theory of Employment, Interest and Money* (1936). Keynes believed that the basic problem of capitalism is not so much its vulnerability to periodic saturations of investment as its likely failure to recover from them. He raised the possibility that a capitalist system could remain indefinitely in a condition of equilibrium despite high unemployment, a possibility not only entirely novel (even Marx believed that the system would recover its momentum after each crisis) but also one made plausible by the persisting unemployment of the 1930s. Keynes therefore raised the prospect that growth would end in stagnation, a condition for which the only remedy he saw was "a somewhat comprehensive socialization of investment."

The quality of growth. A second criticism with respect to market-driven growth focuses on the negative externalities—the adverse side effects—generated by a system of production that is held accountable only to the test of profitability. It is in the nature of a complex industrial society that the production processes of many commodities generate "bads" as well as "goods"—*e.g.*, toxic wastes or unhealthy working conditions, as well as useful outputs.

The externalities of economic activity

The catalog of such market-generated ills is very long. Smith himself warned that the division of labour, by routinizing work, would render workers "as stupid and ignorant as it is possible for a human creature to become," and Marx raised the spectre of alienation as the social price paid for the subordination of production to the imperatives of profit-making. A number of economists have warned that the introduction of technology designed to cut labour costs would create permanent unemployment. In modern times much attention has focused on the power of physical and chemical processes to surpass the carrying capacity of the environment—a concern made cogent by various types of environmental damage arising from excessive discharges of industrial effluents and products. Because these social and ecological challenges spring from the extraordinary powers of technology, they can be viewed as side effects of socialist as well as capitalist growth. But the argument can be made that market growth, by virtue of its overriding obedience to profit, is congenitally blind to such externalities.

Equity. A third criticism of capitalist growth concerns the fairness with which capitalism distributes its expanding wealth or with which it shares its recurrent hardships. This criticism takes on a specific and a general form.

The specific form focuses on disparities in income among layers of the population. At the end of the 20th century in the United States, for example, the lowest fifth of all families received only 3.6 percent of total income, whereas the topmost fifth received 49 percent. Significantly, this disparity results from the concentration of assets in the upper brackets. Also, it is the consequence of highly skewed patterns of corporate rewards that typically give, say, chief executive officers of large companies 50 to 100 times more income than those of ordinary office or factory employees. Income disparities, however, should be understood in perspective, because they stem from a number of causes. As observed in a report from the Federal Reserve Bank of Dallas, "By definition, there will always be a bottom 20 percent, but only in a strict caste society will it contain the same individuals and families year after year."

At a more general level, the criticism may be broadened to an indictment of the market principle itself as the regulator of incomes. The advocate of market-determined distribution declares that in such a society, with certain exceptions, people tend to be paid what they are worth—that is, their incomes reflect the value of their contribution to production. Thus market-based rewards lead to the efficiency of the productive system, thereby maximizing the total income available for distribution. This argument is countered at two levels. Marxist critics contend that labour under capitalism is systematically paid less than its value by virtue of the superior bargaining power of employers, so that the claim of efficiency masks an underlying condition of exploitation. Other critics question the criterion of efficiency itself, one that counts every dollar of input and output but pays no heed to the moral or social or aesthetic qualities of either and that excludes workers from expressing their own preferences as to the most appropriate decisions for their firms.

The conservative defense

Corrective measures. Various measures have been taken by capitalist societies to meet these criticisms, although it must be recognized at the outset that a deep disagreement divides economists with respect to the appropriate corrective measures. A substantial body of conservative thinkers believe that many of the difficulties of the system spring not from its own workings but from well-meaning attempts to block or channel them. Thus, with respect to the problem of instability, supporters of the conservative view believe that capitalism, left as much as possible to itself, will naturally evidence the expansionary thrust that has marked its history and that whatever instabilities appear will tend quickly to correct themselves, provided that government plays a generally passive role. Conservatives do not deny that the system can give rise to qualitative or distributional ills, but they tend to believe that these are more than compensated for by its general expansive properties. Where specific problems remain, such as bad externalities or serious poverty, the conservative prescription often seeks to utilize the market system itself as the corrective agency—*e.g.*, controlling pollution by charging fees on the outflow of wastes rather than by banning the discharge of pollutants, or alleviating poverty through negative income taxes rather than by welfare payments.

Opposing this view is a much more interventionist approach rooted in generally Keynesian and welfare-oriented policies. This view doubts the intrinsic momentum or reliability of capitalist growth and is therefore prepared to use active government means, both fiscal and monetary, to combat recession. It is also more skeptical as to the likelihood of improving the quality or the equity of society by market means and, although not opposing these, looks more favourably on direct regulatory intervention and on specific programs of assistance to disprivileged groups.

Despite this philosophical division of opinion, a fair degree of practical consensus has been reached on a number of issues. Although there are differences in policy style and determination from one nation to the next, all capitalist governments today take measures to overcome recession, whether by lowering taxes, by borrowing and spending, or by easing interest rates, and all pursue the opposite kinds of policies in inflationary times. It cannot be said that these policies have been unqualified successes, either in bringing about vigorous or steady growth or in ridding the system of its inflationary tendencies. Yet, imperfect though they are, these measures have been sufficient to prevent the development of socially destructive depressions or inflations. It is not the eradication but the limitation of instability that has been a signal achievement of all advanced capitalist countries since World War II. It should be noted, however, that these remedial measures have little or no international application. No institution exists to control credit for the world, comparable to the central banks that control it for individual nations; no global spending or taxing authority can speed up, or hold back, the pace of production for the industrial regions as a whole; no agency effectively oversees the availability of credit for the developing nations or the feasibility of the terms on which it may be extended. Thus, some critics of globalization contend that the internationalization of capitalism may exert destabilizing influences for which no policy corrective as yet exists.

Lack of international regulation

A broadly similar appraisal can be made with respect to the redress of specific threats that emerge as by-products from the market system. The issue is largely one of scale. Specific "bads" can often be redressed by market incentives to alter behaviour (paying a fee for returning used bottles) or, when the negative externality is more serious, by outright prohibition (bans on child labour or on dangerous chemical fertilizers). The problem becomes less amenable to control, however, when the market generates externalities of large proportions, such as traffic congestion in cities. The difficulty here is that the correction of such externalities requires the support and cooperation of the public and thereby crosses the line from the economic into the political arena, often making redress more difficult to obtain. On a still larger scale, the remedy for externalities may require international agreements, and these often raise conflicts of interest between the nation generating the ill effects as a by-product of its own production and those suffering from the effects.

A number of remedies have been applied against the distributional problems of capitalism. No advanced capitalist country today allows the market to distribute incomes without supplementing or altering the resulting pattern of rewards through taxes, subsidies, welfare systems, or entitlement payments such as old-age pensions and health benefits. In the United States these transfer payments, as they are called, amount to some 10 percent of total consumer income; in a number of European nations they come to considerably more. The result has been to lessen considerably the incidence of officially measured poverty. Yet these examples of successful government corrective action do not go unchallenged by those economists who are concerned that some of the "cures" applied to social problems may be worse than the "disease." Markets may fail, in other words, but so might governments.

Economic systems may lose some of the decisive differences that have marked them in the past, suggesting instead a continuum on which elements of both market and planning coexist in different proportions. Societies along such a continuum may continue to designate themselves as capitalist and socialist, but they are likely to reveal as many common aspects in the solutions to their economic problems as they may still display important differences.

BIBLIOGRAPHY

General texts: Historical analysis is presented in ROBERT L. HEILBRONER, *The Making of Economic Society*, 8th ed. (1989). An excellent presentation along more functional lines is FREDERIC L. PRYOR, *A Guidebook to the Comparative Study of Economic Systems* (1985). MORRIS BORNSTEIN (ed.), *Comparative Economic Systems: Models and Cases*, 6th ed. (1979), a book of readings, is also recommended.

Socialism: A history of the debate over the economic feasibility of socialism is available in DON LAVOIE, *Rivalry and Central Planning* (1985). A comprehensive reference collection is PETER J. BOETTKE (compiler), *Socialism and the Market: The Socialist Calculation Debate Revisited*, 9 vol. (2000). Other theoretical and historical works include ALEC NOVE, *The Economics of Feasible Socialism* (1983); BRANKO HORVAT, *The Political Economy of Socialism: A Marxist Social Theory* (1982); DAVID L. PRYCHITKO and JAROSLAV VANEK (eds.), *Producer Cooperatives and Labor-Managed Systems*, 2 vol. (1996); and JÁNUS KORNAI, *The Socialist System* (1992). Discussions regarding China, eastern Europe, and the former Soviet Union are in ANDREI SHLEIFER and DANIEL TREISMAN, *Without a Map: Political Tactics and Economic Reform in Russia* (2000); JEFFREY SACHS, *Poland's Jump: The Socialist Calculation Debate Reconsidered to the Market Economy* (1993); and BARRY NAUGHTON, *Growing Out of the Plan: Chinese Economic Reform, 1978–1993* (1995).

The Soviet experience: The classic work on the economic history of the Soviet Union is ALEC NOVE, *An Economic History of the U.S.S.R., 1917–1991*, 3rd ed. (1992). A political history of the Soviet Union is MARTIN MALIA, *The Soviet Tragedy: A History of Socialism in Russia, 1917–1997* (1994, reissued 1996). PETER J. BOETTKE, *Why Perestroika Failed: The Politics and Economics of Socialist Transformation* (1993), discusses systemic problems.

Capitalism: For broad treatments of capitalism, see ADAM SMITH, *An Inquiry into the Nature and Causes of the Wealth of Nations* (1776); and KARL MARX, *Das Kapital*, (vol. 1, 1867); and *Capital: A Critical Analysis of Capitalist Production*, vol. 1 trans. by SAMUEL MOORE and EDWARD AVELING (1886); both works are available in many later editions. ROBERT L. HEILBRONER, *The Nature and Logic of Capitalism* (1985), treats the social formation of capitalism. FERNAND BRAUDEL, *Civilization and Capitalism, 15th–18th Century*, 3 vol. (1982–84; originally published in French, 1979), is a wide-ranging overview. NATHAN ROSENBERG and L.E. BIRDZELL, JR., *How the West Grew Rich: The Economic Transformation of the Industrial World* (1986, reissued 1999), discusses the Industrial Revolution and the rise of capitalism. JOHN KENNETH GALBRAITH, *The New Industrial State*, 4th ed. (1985), is a modern classic. MILTON FRIEDMAN, *Capitalism and Freedom* (1962) and *Free to Choose* (1980), treats economics and public policy from a market-oriented perspective. On the Japanese system, YUTAKA KOSAI and YOSHITARO OGINO, *The Contemporary Japanese Economy*, trans. from Japanese (1984), is useful but technical. (R.L.He./P.J.B.)

Economic Theory

Economic theory is the name commonly given to the more general and abstract parts of economics, the principles. These parts are no less practical than concrete-descriptive or applied economics but are less directly related to immediate problems. The mechanics of price relations or of markets afford a general explanation of the organization of production and distribution insofar as this is actually worked out and controlled through competitive buying and selling—which would largely be true even in a planned or socialistic economy that stopped short of complete military regimentation. This branch of the study bears somewhat the same relation to economic politics that pure physics bears to the engineering sciences. Hence the problems of value and distribution have continued to hold their place among the central concerns of economists. However, there has been a notable—one might say a revolutionary—change in the general character of the analysis. The older classical economists centred their attention on the long-run relations between value and costs and were generally content to dispose of short-run variations of price by merely invoking the formula of supply and demand. This was used without careful analysis of the short-run situation, particularly of the role of demand. Work directed toward filling in this gap had important effects in changing the whole conceptual picture. Enlarged theories of production, distribution, consumption, business fluctuations, and other economic elements have been introduced and continually reconsidered from a variety of viewpoints.

The article is divided into the following sections:

Utility and value 916
  Theories of value 916
    Cost-of-production analysis
    Resource limitations and allocation
  Theories of utility 917
    Marginal utility
    Consumers' surplus
    Utility measurement and ordinal utility
  Prices and incomes 919
    Equilibrium of the consumer
    Changes in prices and incomes
Price 920
  The basic functions of economic systems 920
  The workings of the price system 921
    The choice of occupation
    The conservation of resources
  Limitations and failures of the price system 922
    Private and public price control
    Externalities and the price system
    Imperfect knowledge and tastes
  Noncapitalist price systems 923
Market structure: competition, oligopoly, monopoly 923
  Types of market structures 923
  Market conduct and performance 924
    Pure competition
    Monopolistic competition
    Monopoly
    Oligopoly
  Workable competition 925
Production: the output of the factors of production 926
  Minimization of short-run costs 926
    The production function
    Substitution of factors
    Marginal cost
  Maximization of short-run profits 928
    Marginal cost and price
    Marginal product
  Maximization of long-run profits 929
  Criticisms of the theory 929
Distribution: the shares of the factors of production 929
  Capital and interest 929
    The classical theory of capital
    The Austrian school
    Marginalist and Keynesian theories
    Later thinking
    Interest
  Labour and wages 933
    Classical theories
    Marxian surplus-value theory
    Residual-claimant theory
    Bargaining theory
    Marginal-productivity theory and its critics
    Purchasing-power theory
  Land and rent 935
  General theories of distribution 936
    Aspects of distribution
    Components of the neoclassical, or marginalist, theory
    Criticisms of the neoclassical theory
    Returns to the factors of production
    Dynamic influences on distribution
    Personal income and neoclassical theory
Consumption 939
  Patterns of national consumption 939
  Theories of consumer behaviour 939
    Factors influencing consumers
    Attitudes toward necessities and luxuries
Economic fluctuations: stability and instability 941
  Business cycles 941
    Historical studies of cycles
    Theories of economic fluctuation
  Stabilization theories and policies 944
    Keynesian analysis
    National income accounting
    Monetary policy
    Comparisons of the income and money models
    Interest-rate policy
    The "natural" rate of interest and effective demand
Bibliography 951

Utility and value

Value theory and prices

One purpose of value theory in economics is to explain how the prices of goods and services are determined. This is only a step, however, in the analysis of a deeper problem. The modern industrial economy is characterized by a high degree of interdependence of its parts. The supplier of components or raw materials, for example, must deliver the desired quantities of his products at the right moment and in the desired specifications. In economies such as those of western Europe, North America, and Japan, the coordination of these activities is done through the price system. The relative prices of the various inputs (*e.g.*, labour, materials, machinery) tend to determine the proportions in which they will be used. Prices also affect the relative outputs of the various final products, and they determine who will consume them. Value theory, therefore, studies the structure of these decisions, analyzes the influence of prices, and examines the efficiency of the resulting allocation of resources. Value theory is also applied by business firms and government agencies in their decisions that relate to pricing and the allocation of resources.

THEORIES OF VALUE

Cost-of-production analysis. Modern value theory began with Adam Smith (1776), David Ricardo (1817), and a number of other writers, who are generally lumped together as the classical school. These writers sought to explain pricing primarily on the basis of cost of production. That is, if commodity A costs twice as much to produce as commodity B, the price of A will be pushed toward a level twice as high as that of B. If this were not the case—

if, for example, A sold for three times the price of B—then the greater profitability of investment in A would cause its production to increase and drive down its price, while the production of B would decline, thus raising its price. Prices would finally be driven to the 2:1 ratio of the costs of production.

The classical economists were well aware of the oversimplification in this explanation, but, as with most theoretical analysis, its strength lay in the amount it was able to explain with a very simple model. (It is highly misleading to interpret the classical analysis literally, as a picture of its authors' views of the complex world of reality.) It was soon recognized, however, that the cost-of-production analysis considered only part of the relevant problem. Since cost depends on the quantity produced (*e.g.*, costs per unit may decline as production of an item increases), the analysis must take into account the demand for the product. The analysis of demand was made possible by the theory of utility, developed by H.H. Gossen in Germany (1854), Karl Menger in Austria (1871), Léon Walras in France (1874–77), and W.S. Jevons in England (1871).

The role of utility analysis in value theory will be discussed later. It need only be added at this point that modern value theory, following the lead of the English economist Alfred Marshall (*Principles of Economics*, 8th ed., 1920), considers prices to be determined simultaneously by cost and demand considerations. The analysis also recognizes the complex interdependencies in the system, with demands and supplies of various commodities affecting one another.

The concept of alternative uses

Resource limitations and allocation. The fact that goods have value can be ascribed ultimately to the limitations in the world's material endowment. Man does not have all the arable land, petroleum, or platinum that he would like; their use must be rationed. That is why goods have prices; if they were available in unlimited supply they would be free. Price usually serves as the rationing device whereby their use is kept down to the available supply.

Resources can be said to be scarce in both an absolute and in a relative sense: the surface of the Earth is finite, imposing absolute scarcity; but the scarcity that concerns economists is the relative scarcity of resources in different uses. Materials used for one purpose cannot at the same time be used for other purposes; if the quantity of an input is limited, the increased use of it in one manufacturing process must cause it to become scarcer in other uses.

The cost of a product in terms of money may not measure its true cost to society. The true cost of, say, the construction of a supersonic jet is the value of the schools and refrigerators that will never be built as a result. Every act of production uses up some of society's available resources; it means the foregoing of an opportunity to produce something else. In deciding how to use resources most effectively to satisfy the wants of the community, this opportunity cost must ultimately be taken into account.

In a market economy the relationship between the price of a good and the quantity supplied depends on the cost of making it, and that cost, ultimately, is the cost of *not* making other goods. The market mechanism enforces this relationship. In the first instance, the cost of, say, a pair of shoes is the price of the leather, the labour, the fuel, and other elements used up in producing them. But the price of these inputs, in turn, depends on what they can produce elsewhere—if the handbags that can be produced with the leather are valued very highly by consumers, the price of leather will be bid up correspondingly.

THEORIES OF UTILITY

There are two sides to the analysis of price and value: the supply side and the demand side. If cost can be said to underlie the supply relationship that determines price, the demand side must be taken to reflect consumer tastes and preferences. "Utility" is a concept that has been used to describe these tastes. As already indicated, the cost-of-production analysis of value given above is incomplete, because cost itself depends on the quantity produced. The cost analysis, moreover, applies only to commodities the production of which can be expanded and contracted. The price of a first-folio Shakespeare has no relation to cost of production; it must depend in some sense on its utility to purchasers as it affects their bids.

Marginal utility. The classical economists suggested that this leads to a paradox. They argued that utility could not explain the relative price of fine jade and bread, because the latter was for many consumers essential to life, and hence its utility must surely be greater than that of jade. Yet the price of bread is far lower than that of jade. The theory of marginal utility that flowered toward the end of the 19th century supplied the key to the paradox and provided the basis for today's analysis of demand. Marginal utility was defined as the value to the consumer of an additional unit of some commodity. If, for example, the consumer is offered a choice between 22 and 23 slices of bread for his family, marginal utility measures how much more valuable 23 slices are than 22. It is clear that the magnitude of the marginal utility varies with the magnitude of, say, the smaller of the alternatives. That is, for a family of four, the difference between seven and eight slices of bread per day can be substantial, if the family will still be hungry in either case. But the difference in value between 31 and 32 slices may be negligible. If 31 slices offer enough for everyone to fill his stomach, a 32nd slice may be worth very little. Moreover, the difference in value between 122 and 123 slices may be negative—a 123rd slice may just add to the family's disposal problem. These observations lead directly to the plausible notion that marginal utility in some sense diminishes with the base from which one starts the calculation. With only seven or eight slices the marginal utility (incremental value) of an eighth slice is high. With 31 or 32 slices it is lower, and so on. The less scarce a commodity, the lower is its marginal utility, because its possessor in any case will have enough to satisfy his most pressing uses for it, and an increment in his holdings will only permit him to satisfy, in addition, desires of lower priority.

Marginal utility and price

The consumer will be motivated to adjust his purchases so that the price of each and every good will be approximately equal to its marginal utility (that is, to the amount of money he is willing to pay for an additional unit). If the price of an item is P dollars, for example, and the consumer is considering buying, say, 10 units, at which point the marginal utility of the good to him is M (which is greater than P), the consumer will be better off if he purchases 11 rather than 10 units, since the additional unit costs him P dollars. He will keep revising his purchase plans upward until he reaches the point where the marginal utility of the item falls to P dollars. In sum, the consumer's self-interest will lead him (without conscious calculation) to purchase an amount such that the marginal utility is as close as possible to market price. So long as the consumer selects a bundle of purchases that gives him the most benefit (pleasure, utility) for his money, he must end up with quantities such that the marginal utility of each commodity in the bundle is approximately equal to its price.

It now becomes easy to explain the paradox underlying the relationship between the prices of jade and bread. Because a piece of fine jade is scarce, its marginal utility is high, and consumers are willing to pay comparatively high prices for it. The explanation is perfectly consistent with a utility analysis of demand, so long as one relates price to the marginal utility of the item rather than to its total utility. A family's bread may be very valuable to it, but, if it has enough, the marginal utility of the bread will be small, and this will be reflected in its low price.

The relationship between price and marginal utility is important not because it explains issues like the jade–bread paradox but because it enables one to analyze the relationship between prices and quantities demanded. It also, as a practical matter, permits one to judge how well any portion of the price mechanism is working as a device to secure the efficient satisfaction of the wants of the public, within the limits set by available resources. The conclusion that at any price the consumer will purchase the quantity at which marginal utility is equal to price makes it possible to draw a demand curve showing—to a reasonable degree of approximation—how the amount demanded will vary with price. A curve based on the

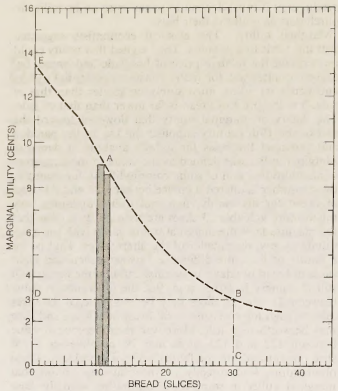


Figure 1: Relationship between marginal utility and quantity (see text).

previous example of bread consumption is given in Figure 1. This shows that if the family gets 10 slices per day the marginal utility of bread will be nine cents (point A). One may reverse the question and ask how much the family would purchase at any particular price, say three cents. The graph indicates that at this price the quantity would be 30 slices, because only at that quantity is marginal utility equal to the three-cent price (point B). Thus the curve in Figure 1, to a reasonable degree of approximation, may be able to do double duty: it may serve as a marginal-utility curve relating marginal utility to quantity and, at the same time, as a demand curve relating quantity demanded to price.

Consumers' surplus. Figure 1 leads to an important conclusion about the consumer's gains from his purchases. The diagram shows that the difference between 10 and 11 slices of bread is worth nine cents to the consumer (marginal utility = nine cents). Similarly, a 12th slice of bread is worth eight cents (see the shaded bars). Thus, the two slices of bread together are worth 17 cents, the area of the two rectangles together. Suppose the price of bread is actually three cents, and the consumer, therefore, purchases 30 slices per day. The total value of his purchases to him is the sum of the areas of all such rectangles for each of the 30 slices; *i.e.*, it is (approximately) equal to all of the area under the demand curve; that is, the area defined by the points 0CBE. The amount the consumer pays, however, is less than this area. His total expenditure is given by the area of rectangle 0CBD—90 cents. The difference between these two areas, the quasi-triangular area DBE, represents how much more the consumer would be willing to spend on the bread over and above the 90 cents he actually pays for it, if he were forced to do so. It represents the absolute maximum that could be extracted from the consumer for the bread by an unscrupulous merchant who had cornered the market. Since, normally, the consumer only pays quantity 0CBD, the area DBE is a net gain derived by the consumer from the transaction. It is called consumers' surplus. Virtually every purchase yields such a surplus to the buyer.

The concept of consumers' surplus is important for public policy, because it offers at least a crude measure of the public benefits of various types of economic activity. In deciding whether a government agency should build a dam, for example, one may estimate the consumers' surplus from the electricity the dam would generate and seek to compare it with the surplus that could be yielded by alternative uses of the resources needed to construct and operate the dam.

Consumers' preferences

Utility measurement and ordinal utility. As originally conceived, utility was taken to be a subjective measure of strength of feeling. An item that might be described as worth "40 utils" was to be interpreted to yield "twice as much pleasure" as one valued at 20 utils. It was not long before the usefulness of this concept was questioned. It was criticized for its subjectivity and the difficulty (if not impossibility) of quantifying it. An alternative line of analysis developed that was able to accomplish most of the same purposes but without as many assumptions. First introduced by the economists F.Y. Edgeworth in England (1881) and Vilfredo Pareto in Italy (1896–97), it was brought to fruition by Eugen Slutsky in Russia (1915) and J.R. Hicks and R.D.G. Allen in Great Britain (1934). The idea was that to analyze consumer choice between, say, two bundles of commodities, A and B, given their costs, one need know only that one is preferred to another. This may at first seem a trivial observation, but it is not as simple as it sounds.

Indifference analysis

In the following discussion, it is assumed for simplicity that there are only two commodities in the world. Figure 2 is a graph in which the axes measure the quantities of two commodities, X and Y. Thus, point A represents a bundle composed of seven units of commodity X and five units of commodity Y. The assumption is made that the consumer prefers to own more of either or both commodities. That means he must prefer bundle C to bundle A, because C lies directly to the right of A and hence contains more of X and no less of Y. Similarly, B must be preferred to A. But one cannot say, in general, whether

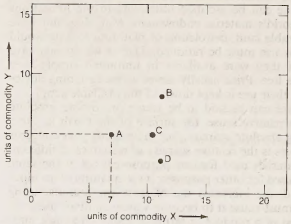


Figure 2: Commodities X and Y (see text).

A is preferred to D or vice versa, since one offers more of X and the other more of Y.

The consumer may in fact not care whether he receives A or D—that is, he may be indifferent (see Figure 3). Assuming that there is some continuity in his preferences, there will be a locus connecting A and D, any point on which (E or A or D) represents bundles of commodities of equal interest to this consumer. This locus (I–I′ in Figure 3) is called an indifference curve. It represents the consu-

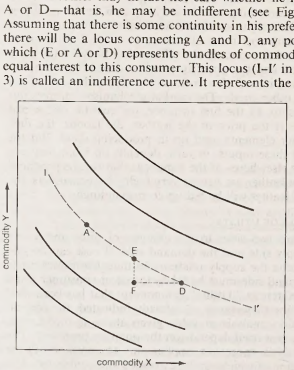


Figure 3: Indifference curves (see text).

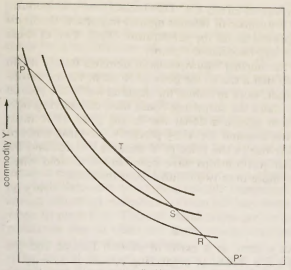


Figure 4: Indifference curves and a price line (see text).

mer's subjective trade off between the two commodities—how much more of one he will have to get to make up for the loss of a given amount of another. That is, one may treat the choice between bundle D and bundle E as involving the comparison of the gain of quantity FD of X with the loss of FE of Y. If the consumer is indifferent between D and E, the gain and loss just offset one another; hence, they indicate the proportion in which he is willing to exchange the two commodities. In mathematical terms, FE divided by FD represents the average slope of the indifference curve over arc ED; it is called the marginal rate of substitution between X and Y.

Figure 3 also contains other indifference curves, some representing combinations preferred to A (curves lying above and to the right of A) and some representing combinations to which A is preferred. These are like contour lines on a map, each such line being a locus of combinations that the consumer considers equally desirable. Conceptually, through every point in the diagram there is an indifference curve. Figure 3, with its family of indifference curves, is called an indifference map. This map obviously does no more than rank the available possibilities; it indicates whether one point is preferred to another but not by how much it is preferred.

It is easy to show that at any point such as E the slope of the indifference curve, roughly FE divided by ED, equals the ratio of the marginal utility of X to the marginal utility of Y for the corresponding quantities. For in moving from E to D the consumer gives up FE of Y, a loss valued, by definition, at approximately FE multiplied by the marginal utility of Y, and he gains FD of X, a gain worth FD multiplied by the marginal utility of X. Relative marginal utilities can be measured in this way because their ratio does not measure subjective quantities—rather, it represents a rate of exchange of two commodities. The marginal utility of X measured in money terms tells one how much of the commodity used as money the consumer is willing to give for more of the commodity X but not what psychic pleasure the consumer gains.

PRICES AND INCOMES

Indifference curves and the price line

One other type of information is needed to complete the analysis of consumer choice: the prices of X and Y and the amount the consumer has available to spend. In what follows, it will be assumed that the consumer spends all his money on the available commodities (savings bonds being among the commodities). If P_X and P_Y are the prices of commodities X and Y, respectively, and M represents the amount of money available for spending, the condition that all of the money is spent yields the equation

$$P_X X + P_Y Y = M \qquad (1)$$

or, solving for Y in terms of X,

$$Y = -\frac{P_X}{P_Y} X + \frac{M}{P_Y}. \qquad (2)$$

This is obviously the equation of a straight line with slope $-\frac{P_X}{P_Y}$ and with y-intercept $\frac{M}{P_Y}$. The line, called the budget line, or price line, represents all the combinations of X and Y that the consumer can afford to buy with income M at the given prices.

Equilibrium of the consumer. Figure 4 combines this price line and the indifference curves, permitting direct analysis of the consumer purchase decision. Line PP′ is the price line corresponding to equation (2) above. Any point R on that line represents a combination of X and Y that a given consumer can afford to purchase; however, R is not an optimal choice. This can be seen by comparing R with S on the same price line. Since S lies on a higher indifference curve than R, the former is the preferred position, and, since S costs no more than R (they are on the same price line, so each costs M dollars), S gives the consumer more for his money. It is at T_1 however, the point of tangency between the price line and an indifference curve, that the consumer reaches his highest indifference curve; this is, therefore, the optimal point for him, given his pattern of tastes as shown by the shapes of his indifference curves. This is the solution of the choice problem—it explains, in principle, the consumer's purchase decision on the basis of his given preferences, with no assumptions as to degrees of measurable utility.

The tangency at the solution point has a significant interpretation. It was noted above that the slope of the indifference curve is the ratio of the marginal utilities of the two commodities. It follows that, at the optimal point T_1 a dollar of expenditure must offer the same utility whether spent on X or on Y. If this is not so—as at point R in Figure 4, where the consumer gets more for his money by spending a dollar on Y rather than on X—it will pay him to reallocate his expenditures between the two commodities accordingly, moving toward S from R.

Changes in prices and incomes. The diagram becomes more illuminating when one investigates how the consumer's decision is affected by a change in his income or in the price of a commodity. Equation (2) indicates that a change in income, M, does not affect the slope of the price line, only its intercept. Thus, as the person's income increases, the price line undergoes a sequence of parallel shifts (Figure 5). For each such line there will be a point of tangency, T_i with an indifference curve, showing the consumer's optimal bundle of purchases with the corresponding income. The locus of these points (T_1, T_2, T_3 . . .) may be called the income–consumption curve; it shows how the consumer's purchases vary with his income. Normally the curve will have a positive slope, as EE′ does in Figure 5A, meaning that as a person grows wealthier he will buy more of each commodity. But the slope can be negative for some stretches (GG′ in Figure 5B). In that case, X is said to be an inferior good of which the consumer buys less as his income rises.

Income–consumption curve

The diagram can also be used to show what happens as the price of X varies. From equation (2) it can be seen that the Y-intercept is not affected by an increase in the price of X but that the slope of the price line grows. Thus, as P_X rises, the price line shifts from PP′ to PR′ in Figure 6. This means that, as P_X rises, M dollars will buy as much of good Y as before (the position of point P at which all

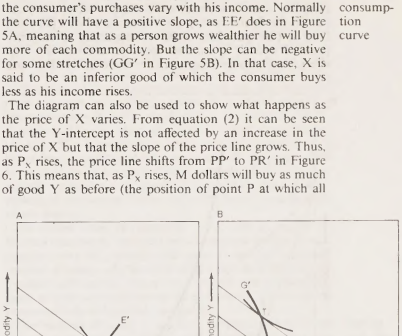


Figure 5: (A) Positive and (B) negative income–consumption curves (see text).

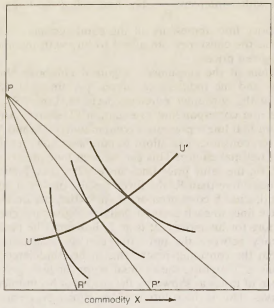


Figure 6: Price–consumption curve (see text).

M dollars are spent on commodity Y does not change), but that M dollars will now buy less of good X, so that the position of point P′ must move toward the left. Once again, by following the points of tangency between indifference curves and the price lines for various values of P_X, one contains a locus UU′, the price–consumption curve, showing how the consumer's purchases vary with P_X.

It is useful to divide the effects of the price change conceptually into two parts. An increase in the price of X obviously affects the relative cost of X and Y. But it also decreases the consumer's overall purchasing power. The effect on purchases of this reduction of purchasing power is called the income effect of the price change. Its effect via the relative price change is called the substitution effect. The division can be carried out graphically as follows: let the price of X increase so that the price line in Figure 7 moves from PP′ to PR′, and assume an imaginary intermediate price line, LL′, with the slope of PR′ but tangent to the indifference curve that was attained with the old price line PP′. The imaginary price line has the following properties: (1) it involves the same real income as PP′ (tangency points T and S are the same indifference curve), and (2) it involves the same relative prices as the new price line since their slopes are the same. The rise in price has, in the figure, caused the demand for X to fall from C to A (the quantities of X corresponding to tangency points T and U). It has been possible to divide the total effect, CA, into two parts, the income effect, BA, and the substitution effect, CB. This breakdown is important, because a number of interesting and important theorems can be proved about the substitution effect. Two of these theorems will illustrate the point.

Income and substitution effects

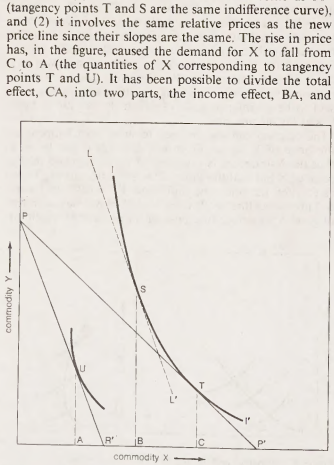


Figure 7: Income effect and substitution effect (see text).

Under the normal assumptions of demand theory it can be proved that a rise in the price of X must, via the substitution effect, work to reduce the demand for X; the second theorem states the surprising result that, considering only substitution effects, a dollar rise in the price of X must change the demand for Y by precisely the same amount as a dollar rise in the price of Y changes the demand for X. Similar relationships have been shown to hold when there are more than two commodities involved.

(W.J.B.)

Price

The price system, as it exists in western Europe and the Americas, is a means of organizing economic activity. It does this primarily by coordinating the decisions of consumers, producers, and owners of productive resources. Millions of economic agents who have no direct communication with each other are led by the price system to supply each other's wants. In a modern economy the price system enables a consumer to buy a product he has never previously purchased, produced by a firm of whose existence he is unaware, which is operating with funds partially obtained from his own savings.

Price as consensus value

Prices are an expression of the consensus on the values of different things, and every society that permits exchanges among men has prices. Because prices are expressed in terms of a widely acceptable commodity, they permit a ready comparison of the comparative values of various commodities—if shoes are $15 per pair and bread 30 cents per loaf, a pair of shoes is worth 50 loaves of bread. The price of anything is its value in exchange for a commodity of wide acceptability (money): the price of an automobile may be some 50 ounces of gold or 25 pieces of paper bearing the picture of an eminent statesman.

A system of prices exists because individual prices are related to each other. If, for example, copper rods cost 40 cents a pound and the process of drawing a rod into wire costs 25 cents a pound, then, if the price of wire exceeds 65 cents, it will be profitable to produce wire; and if the price of wire falls below 65 cents, it will be ruinous to produce wire. Competition, therefore, will hold the price of wire about 25 cents per pound above that of rods. A variety of such economic forces ties the entire structure of prices together.

The system of prices can be arranged to reward or penalize any kind of activity. Society discourages the production of electric shoestring-tying machines by the simple expedient of making such a machine's attainable selling price less than the prices of the resources necessary to produce it. Society stimulates people of large athletic promise to learn golf (rather than polo or cricket) by the immense prizes (= prices) that are given to tournament winners. The air in many cities is dirty because no one is charged a price for dirtying it and no one can pay a price for having it cleaned.

THE BASIC FUNCTIONS OF ECONOMIC SYSTEMS

Every economic system has three functions. In a decentralized (usually private enterprise) economic system, the price mechanism is the instrument by which these functions are performed.

One function is to determine what is to be produced and in what quantity. Even a primitive economy must choose between food and shelter, weapons and tools, priests and hunters. In a modern economy the potential variety of goods and services that may be produced is immense. Consider simply the 10,000 new book titles that are published each year or the hundreds of colours of paint or the thousands of styles of clothing that are produced—each of these actual collections being much smaller than modern technology permits.

A price system weighs the desires of consumers in terms of the prices they are willing to pay for various quantities of each commodity or service. The payment of $1,000 for

the services of a skilled surgeon (a price much influenced by the number of surgeons) reflects the great importance of his services to the buyer-patient, whereas the offer of 75 cents for a month's use of an additional telephone outlet reflects the minor convenience it provides. Of course the offers of consumers are influenced by their wealth as well as their desires, but for any one consumer relative desires are proportional to price offers.

Universal laws are most uncommon in social life. Economists nonetheless place immense confidence in the proposition that the consumer will buy less of any commodity when its price rises. This *law of demand* is by no means a necessary fact of life; rather it is an empirical rule to which there are no known, reliable exceptions. Bread, caviar, education, narcotics—a man will buy more of each when its price falls. These demand prices are the guides to producers that in effect tell them which commodities to produce and in what quantities.

The second function an economy must perform is to decide how the desired goods are to be produced. There is more than one way not only to skin a cat but also to grow wheat, train lawyers, refine petroleum, and transport baggage. The efficient production of goods requires that certain obvious rules be followed: no resource should be used in producing one thing when it could be producing something more valuable elsewhere; and each product should be made with the smallest possible amount of resources.

Allocation of resources

A functioning price system steers resources into their most important use by appealing to the desires of their owners for large incomes. For example, the person capable of being a surgeon is drawn to this occupation from, say, that of a high school teacher by the promise of annual earnings (= price of labour) much more than those of the high school teacher. Capital is drawn from a faltering trade to a booming new industry in which it receives a higher return.

This same price system seeks to achieve production efficiency through the sanction of competition. If one firm, for instance, can produce shoes with fewer resources than its rivals, it will make larger profits; so it is stirred to discover more efficient combinations of inputs and location of plant, to devise wage systems to stimulate its workers, to use computers to reduce inventories, and so forth without end.

The third function of an economy is to determine who gets the product. Family A gets $5,000 worth of goods this year, family B five times as much—how is the division to be decided? The incomes of individuals are determined by the quantities of resources (labour skills, capital in all its forms) they own and the prices they receive for the use of these resources. Workers are incited by the price system to acquire new skills and to exercise them diligently, and families are encouraged to savings (capital accumulation) by the payment of interest or dividends. The inheritance of both personal ability and wealth also enter into the distribution of income.

If the price system is working reasonably well (some of the common failures will be noted later), it performs all of these economic functions with remarkable subtlety and precision. Society desires not only the correct amount of wheat but also that it be consumed more or less evenly over the crop year, with a surplus to carry over in case of a partial failure of the next year's crop. The price system provides a seasonal price pattern that encourages the holding of inventories rather than early splurging and richly rewards speculators who correctly anticipate a crop failure and hold grain that will alleviate it. In the same way, the desires of every sizable group of consumers (or resource owners) are registered through the price system; entrepreneurs are incited by price offers to provide opera and musical comedy, kosher food, and Persian delicacies. One might almost say that the price system is devoted to minority rule, since the only pressure toward uniformity is in the possibility of lowering costs of production by standardizing goods.

High prices in a properly functioning price system thus serve as incentives to produce more and consume less, and lower prices serve as corresponding deterrents. In addition the price system is a method of communicating information. Herbert Spencer once stated, rather ponderously, that only by constant iteration can alien truths be impressed upon reluctant minds: the price system, with its capacity for infinite repetition, is well suited to this sometimes unpleasant task. A higher price of steel scrap, for example, tells thousands of owners and collectors of scrap that more scrap is wanted and that more exhaustive search for abandoned rails, boilers, radiators, and machines is worth undertaking. A higher price of gasoline tells thousands of automobile drivers that gasoline should be used more sparingly, and the message is repeated each time each driver purchases more gasoline.

THE WORKINGS OF THE PRICE SYSTEM

The complexity and variety of tasks performed by the price system will be illuminated by an examination of three specific economic problems.

The choice of occupation. Individuals must be distributed among occupations in such a way as to serve two basic purposes. First, the labourer must be placed where he is most productive—making certain that Enrico Fermi becomes a physicist rather than a chef and that there are not too many plumbers and too few electricians. Second, the individual worker should be given an occupation that is congenial to him: since he will spend a large part of his life at work, it will be a better life if he can choose the type he prefers.

The price of labour

The price of labour is the instrument by which workers are distributed among occupations: wages in rapidly growing occupations and rapidly growing parts of the nation are higher than in corresponding employment in declining occupations and areas. The choice of occupation involves, however, much more than simply a comparison of wage rates. The following are a few of the complications: (1) The wages of an occupation must as a rule be sufficient to compensate the costs of training. (2) The wages of an occupation must be sufficient to compensate special disadvantages (such as a large chance of unemployment). (3) Wages must be higher in large cities than in small because living costs are higher in large cities. (4) Wages must compensate workers for their additional skill as they acquire experience (they usually reach a peak of earnings between ages 40 and 55) and thereafter decline as the worker's efficiency declines. (5) Wages will reflect differences in taxation, fringe benefits (pensions, vacations), etc. Accordingly the wage structure even for a single occupation in a single city is elaborate. When a single wage (price) is imposed upon an occupation, labourers are no longer properly distributed by wages; for example, a city school system that pays all teachers of given experience the same wage finds it difficult to staff its less attractive schools.

The preferences of the individual worker cannot be given full play, or each person would become president of the corporation at a sumptuous salary. Yet the labourer may choose to live in California rather than Maine; then the price system will incite employers to move their operations to California, where they can hire this labourer more cheaply. The labourer may prefer to work long hours or short hours, and employers are induced by wage offers to cater to the labourer's diverse preferences. In fact it is equally appropriate to speak of the worker buying conditions of work and of the employer buying the services of the worker.

The conservation of resources. A society has some resources that can be replaced by investment: timber is now largely grown as a commercial crop. Farmland is a more ancient example: the fertility of soil can be increased by prudent cultivation. Other resources are not replaceable, such as coal and petroleum. How does the price system conserve these exhaustible resources?

Pricing and conservation

The method of using a resource is independent of the pattern over time of income and expenditures that the owner of the resource desires. Suppose that a farm will have a value of $100,000 if it is maintained at a constant level of fertility and yields a yearly income of $10,000 forever but that it can be cultivated ("mined") intensively to yield $12,000 a year for five years at the cost of a much reduced yield thereafter, with a value of $90,000. Even if the farmer is in urgent need of immediate funds

and does not expect to live more than five years, he will still cultivate the farm at the uniform rate. Only then is it worth its maximum value to him, and only then (by sale or mortgage) can he obtain the largest possible funds even in the near future. In short, one need not adapt his expenditure pattern to his income pattern so long as he can borrow or lend.

If the growth of consumption or the decline of reserves threaten the exhaustion of supplies of a resource, then the price of that resource will rise and promise to rise more in the future, and this rise will serve to reduce current consumption and to reward the owner of the resource for holding back much of the supply for the future. This rise in price will therefore also stimulate buyers to find more economical ways of using the commodity (for example, burning the fuel more efficiently) and stimulate producers to find new supplies or substitute products. The price system will, therefore, ensure that the supply of the resource will be stretched out so that the resource will be available in both the present and the future.

LIMITATIONS AND FAILURES OF THE PRICE SYSTEM

The price system is an extraordinarily powerful instrument in organizing an economic system, but it is subject to three broad classes of limitations.

Private and public price control. Sometimes prices are not permitted to do their work. Monopolies are able to exert control over prices; and they use it, sensibly enough, to raise their profits above the level allowed by competition. The monopolist (or group of colluding enterprises) sets prices at a level such that prices are above costs or, to use words of identical significance, such that resources earn more in the monopolized industry than they can earn elsewhere. The basis of the monopoly is its ability to prevent outsiders from entering the industry to share in the unusual profits and, by the act of producing, actually serve to eliminate them.

The cost of monopoly

The fixing of prices by monopolists reduces the income of society. This is, in fact, the only well-established criticism (on grounds of efficiency) to be levied against monopolies: there is no reason to assume that they will make products less suited to consumer tastes or innovate more slowly or pay lower wages or otherwise misallocate resources. But the basic inefficiency led, first in the United States in 1890 and then increasingly in European nations, to governmental policies to maintain or restore competition.

Public price control has two aspects. A large part of public regulation is intended to correct monopolistic pricing (or other failures of the price system); this includes most public-utility regulation in the United States (transportation, electricity, and gas, etc.). Whatever the success of these endeavours—and on the whole there has been a substantial decline in confidence in the regulatory bodies—they are usually instructed to achieve the goals of an efficient price system.

Other public price controls are designed to serve ends outside the reach of the price system. Prices of farm products are regulated (raised) in most nations with the intention of improving farmers' incomes, and the fixing of interest rates paid by banks is undertaken to improve bank earnings. Such policies are invariably defended on various economic and ethical grounds but reflect primarily the political strength of large and well organized producer groups.

Externalities and the price system. A second class of limitations consists of things that should be done but are not performable by a price system. Even when prices are freely established by competition, there is a class of economic relationships called "externalities" not efficiently controlled by prices. These may be illustrated by the air pollution caused by automobiles. Since no single automobile makes a significant contribution to air pollution, the owner has no incentive to bear the cost of installing antipollution devices even though all drivers would be better off if each did so. Yet if there are many automobiles in a region, it would be prohibitively expensive for drivers to contract with one another to have each install devices in his automobile to reduce pollution. The external effects of any one automobile's exhaust fumes are so diffuse and affect any one person so triflingly that they cannot be regulated by the price system.

The class of "externalities" is as broad as the class of actions that have effects upon people who are not parties to the contracts governing the actions. An attractive garden pleases passers-by, but they cannot be charged a portion of its cost. A new piece of scientific knowledge will prove useful to unknown persons. These two examples indicate that some externalities are economically trivial and some are highly important.

When the price system cannot deal with diffused effects, other social controls often take its place. The state invokes a whole arsenal of policies to deal with externalities, of which the following are only examples: (1) The state may subsidize activities that do not end in a product that can be sold. Thus basic scientific research that does not lead to patentable processes is subsidized. (2) Individuals may be compelled to act uniformly in areas where contracts would be too expensive; traffic laws, zoning laws, and compulsory vaccination are examples. (3) The state may itself undertake an activity that cannot be financed by sale of services, the most obvious example being national defense.

Traffic congestion and prices

An interesting type of externality is the problem of highway congestion. Any one person's presence on a highway at a time and place of peak density has only a negligible effect upon others, so that, except on toll roads, private contracts have not been feasible. The state itself has not been able to deal effectively with highway congestion. More highways can be built until no highway is ever crowded, but this would be intolerably expensive. The state has lacked a method of inducing drivers to shift to less crowded hours and routes by charging fees to those drivers who impose high congestion costs by driving at peak times. Recent developments in technology may make it feasible to use the price system to reduce congestion. For example, cameras at appropriate points could photograph automobile licenses and a computer could accumulate the charges based on route and time for each automobile. Then only a person for whom travel at peak times was worth, for example, 25 cents per mile would impose (and pay for) the congestion he created.

Imperfect knowledge and tastes. A third class of limitations to which the price system is subject has to do with the control of knowledge and tastes. To the extent that an economic actor, whether a consumer, a labourer, or an investor, is poorly informed, he is likely to make decisions whose consequences are much different from those he desired and expected. What follows relates only to consumer decisions, but parallel issues arise in labour markets, securities markets, etc.

A consumer can satisfy his desires only if he makes intelligent purchases—that is, only if the goods he buys are what he believes them to be. How can the consumer know whether the meat is free of disease or whether the washing machine will function well and long or whether the fabric of the garment is one synthetic fibre or another? To ascertain these facts personally, the consumer would have to be a versatile scientist equipped with a superb laboratory—and then he would need to spend so much time testing goods that he would have little time to enjoy them.

In some measure the consumer does experiment in his buying: whenever he buys a thing repeatedly, experience tells him much concerning its properties. Direct experience is a sufficient guide in buying celery or hiring domestic servants, but usually the purchase of information takes a less direct form. The city's premier department store can sell at prices somewhat higher than less well-known retailers; and the difference represents the payment of a price for reliability, responsibility, and the guarantee of quality. In parallel fashion the consumer buys the washing machine of a company that made his excellent refrigerator. Occasionally, information is bought directly: the advice of a lawyer, the knowledge of an appraiser, the taste of an interior decorator.

The need for information

The most important and controversial method of informing consumers is by advertising. Many critics are outraged by the self-serving statements of sellers, some of whom indubitably provide irrelevance and deception rather than information. Yet the informational content of advertising

may not be as deficient as its critics believe: advertising itself meets two market tests. In the first place, the direct sale of information by consumer advisory services has never become important, although there are no obstacles to entering this business. In the second place, there has been a general, sustained improvement in the quality of consumer goods over time: the automobile tire goes many more miles than formerly; the airplane flies more safely.

Nevertheless, recent public policy has paid great attention to increasing the safety of products and to raising the accuracy of advertising claims.

Knowledge is sometimes difficult to distinguish from taste: does the consumer who persists in smoking cigarettes have inadequate knowledge or simply different comparative values for the pleasures and risks of smoking? Censorship, in any event, is fairly common in every economic system: no society allows young children or incompetents full freedom of action or allows the unlimited sale of narcotics. Since the price system never forbids an effective demand (a demand backed by a willingness to pay the supply price), some form of restriction of prices is, therefore, necessary if certain tastes are to be forbidden or restricted. Compulsory school attendance can be viewed as, in effect, a form of censorship; and so are the controls on sale of firearms and the taxes on tobacco and liquor.

NONCAPITALIST PRICE SYSTEMS

The foregoing discussion has been confined to the price system as it exists in capitalist economies. The Communist countries have prices, but not autonomous price systems; in those countries the direction of economic activity is largely in the hands of the central authorities, and prices are used mainly as a marketing device. None of the three allocative functions of an economy—determination of what will be produced, of how it will be produced, and of who will get the product—is performed by the price mechanism in the socialist economies.

"Shadow prices"

The relative scarcities that money prices measure exist, of course, in all countries and would exist in a world where no money or exchanges were allowed. Robinson Crusoe had a problem of allocating his time between sleep, garnering food, building shelter, etc.; and he confronted implicit costs of extending any one activity, for more food meant less of other things. The economist calls these implicit exchange ratios "shadow prices," and they appear in all areas of life in which deliberate choices are made.

Price systems are therefore the result of scarcity. The basic proposition of economics, that scarcities are essentially ubiquitous, is often phrased as "there is no such thing as a free lunch"; and it reminds one that the price of the lunch may be future patronage, a reciprocal lunch, or a boring monologue. The task of economic organization is the task of devising price systems that allow a society to achieve its basic goals. (G.J.S.)

Market structure: competition, oligopoly, monopoly

When economists use such words as "competition" and "monopoly" they have in mind certain complex relations among firms in an industry. An industry, as economists define it, is a group of sellers of close-substitute products who supply a common group of buyers. For the economy as a whole an industry would include all sellers having this relation. Thus one can recognize a cigarette or automobile or aluminum industry—in all, hundreds of industries.

TYPES OF MARKET STRUCTURES

Different industries have different market structures—that is, different market characteristics that determine the relations of sellers to one another, of sellers to buyers, and so forth. Probably the most important aspects of market structure are (1) the degree of concentration of sellers in an industry, (2) the degree of product differentiation, and (3) the ease or difficulty with which new sellers can enter the industry.

Seller concentration refers to the number of sellers in an industry together with their comparative shares of industry sales. When the number of sellers is quite large, and each seller's share of the market is so small that in practice he cannot, by changing his selling price or output, perceptibly influence the market share or income of any competing seller, economists speak of *atomistic competition.* A more common situation is that of *oligopoly,* in which the number of sellers is so few that the market share of each is large enough for even a modest change in price or output by one seller to have a perceptible effect on the market shares or incomes of rival sellers, and to cause them to react to the change. In a broader sense, oligopoly exists in any industry in which at least some sellers have large shares of the market, even though there may be an additional number of small sellers. When a single seller supplies the entire output of an industry, and thus can determine his selling price and output without concern for the reactions of rival sellers, a *single-firm monopoly* exists.

Product differentiation

The structure of a market is also affected by the extent to which those who buy from it prefer some products to others. In some industries the products are regarded as identical by their buyers—as, for example, basic farm crops. In others the products are differentiated in some way so that various buyers prefer various products. Notably, the criterion is a subjective one; the buyers' preferences may have little to do with tangible differences in the products but are related to advertising, brand names, and distinctive designs. The degree of product differentiation as registered in the strength of buyer preferences ranges from slight to fairly large, tending to be greatest among infrequently purchased consumer goods and "prestige goods," particularly those purchased as gifts.

Industries vary in the ease with which new sellers can enter them. The barriers to entry consist of the advantages that sellers already established in an industry have over the potential entrant. Such a barrier is generally measurable by the extent to which established sellers can persistently elevate their selling prices above minimal average costs without attracting new sellers. The barriers may exist because costs for established sellers are lower than they would be for new entrants, or because the established sellers can command higher prices from buyers who prefer their products to those of potential entrants. The economics of the industry also may be such that new entrants would have to be able to command a substantial share of the market before they could operate profitably.

The effective height of these barriers varies. One may distinguish three rough degrees of difficulty in entering an industry: *blockaded entry,* which allows established sellers to set monopolistic prices, if they wish, without attracting entry; *impeded entry,* which allows established sellers to raise their selling prices above minimal average costs without attracting new sellers, but not as high as a monopolist's price; and *easy entry,* which does not permit established sellers to raise their prices at all above minimal average costs without attracting new entrants.

This discussion of market characteristics suggests a general way of classifying industries according to their market structures:

I. Atomistic competition
  A. With homogeneous products—"pure competition"
  B. With differentiated products—"monopolistic competition"
II. Oligopoly
  A. With homogeneous products—"pure oligopoly"
    1. With blockaded entry
    2. With impeded entry
    3. With easy entry
  B. With differentiated products—"heterogeneous oligopoly"
    1. With blockaded entry
    2. With impeded entry
    3. With easy entry
III. Single-firm monopoly
  A. With blockaded entry
  B. With impeded entry

Under atomistic competition, in which entry is generally easy, there are no barriers to entry. By the same token, product differentiation among sellers is obviously inconsistent with single-firm monopoly.

The comparative importance of these types of market structure differs among various sectors of the economy.

In the manufacturing sector in the United States, which includes about 400 industries, single-firm monopolies are almost completely absent. But in more than half of the manufacturing industries there is enough seller concentration for them to qualify as oligopolies. The remaining industries are more or less atomistic in their market structure. An appreciable degree of product differentiation is found in about half of the oligopolistic industries and in about half of the atomistic industries. Very strong product differentiation is usually found among oligopolistic industries. Easy entry is typical of atomistic industries, impeded entry of oligopolies. Entry is probably blockaded in a minor fraction of the latter, generally those with very high seller concentration.

The proportions of oligopolistic and atomistic manufacturing industries are about the same in the United Kingdom as in the United States. The incidence of oligopolies is slightly higher in Japan and progressively higher in France, Italy, Canada, and Sweden. Single-firm monopolies in manufacturing are found in a few industries in some of these countries, but they are typically under government ownership.

In the public-utility sector in the United States, single-firm monopolies are typically found in industries supplying gas, electricity, and telephone and telegraph service. Oligopoly, frequently highly concentrated, is typical in the radio and television and transportation industries; entry into these industries is usually very difficult or blockaded. The significance of such structural conditions is lessened, however, by the fact that these industries are subject to various degrees of public regulation. The situation is much the same in other Western countries, except that public utilities are frequently under government ownership.

In the distributive trades (wholesaling and retailing), a number of United States industries are fairly atomistic, while a somewhat larger number are relatively unconcentrated oligopolies in which a few large sellers supply about half the industry's output and a very large number of small sellers supply the remainder. Product differentiation is important. Entry is relatively easy. The service-trade industries in the United States display a similar range of characteristics. In the distributive- and service-trade sectors in other Western countries, oligopoly is less frequent and atomistic industries are proportionally more important. The residential-construction industries in the U.S. and elsewhere are relatively atomistic in structure, have significant product differentiation, and are easy to enter. Industries in the agricultural sectors of Western countries generally are typically atomistic in structure, with easy entry. But the significance of these structural conditions is lessened by governmental interference designed to modify the working of market forces.

MARKET CONDUCT AND PERFORMANCE

Sellers' behaviour in different market structures

How do sellers behave in determining their selling prices, outputs, advertising costs, and so forth; and in what ways does this market behaviour differ among industries with different types of market structure? An educated layman might ask, for example, whether sellers cut their selling prices in order to take customers away from each other until some rock-bottom market price is reached just high enough to allow them minimal interest returns on their investments or whether, on the other hand, they agree with each other to set a uniform higher price well above their production costs, sharing the market and reaping excessive profits.

It is helpful to distinguish the related ideas of *market conduct* and *market performance*. Market conduct refers to the price and other market policies pursued by sellers, in terms both of their aims and of the way in which they coordinate their decisions and make them mutually compatible. Market performance refers to the end results of these policies—the relationship of selling price to costs, the size of output, the efficiency of production, progressiveness in techniques and products, and so forth.

Pure competition. Market conduct and performance in atomistic industries provide good standards against which to measure behaviour in other types of industry. The atomistic category includes both pure competition and monopolistic competition. In pure competition, a large number of small sellers supply a homogeneous product to a common buying market. In this situation no individual seller can perceptibly influence the market price at which he sells but must accept a market price that is impersonally determined by the total supply of the product offered by all sellers and the total demand for the product of all buyers. The large number of sellers precludes the possibility of a common agreement among them, and each must therefore act independently. At any going market price, each seller tends to adjust his output to that quantity that will yield him the largest aggregate profit, assuming that the market price will not change as a result. But the collective effect of such adjustments by all sellers will cause the total supply in the market to change significantly so that the market price falls or rises. Theoretically, the process will go on until a market price is reached at which the total output that sellers wish to produce is equal to the total output that all buyers wish to purchase. This way of reaching a provisional equilibrium price is what Adam Smith was referring to when he wrote of prices being determined by "the invisible hand" of the market.

The competitive ideal

If the provisional equilibrium price is high enough to allow the established sellers profits in excess of a normal interest return on investment, then added sellers will be drawn to enter the industry, and supply will increase until a final equilibrium price is reached that is equal to the minimal average cost of production (including an interest return) of all sellers. Conversely, if the provisional equilibrium price is so low that established sellers incur losses, some will tend to withdraw from the industry, and supply will decline until the same sort of long-run equilibrium price is reached.

The long-run performance of a purely competitive industry therefore embodies these features: (1) industry output is at a feasible maximum and industry selling price at a feasible minimum; (2) all production is undertaken at minimum attainable average costs, since competition forces them down; and (3) income distribution is not influenced by the receipt of any excess profits by sellers.

This performance has often been applauded as ideal from the standpoint of general economic welfare. But the applause, for several reasons, should not be unqualified. Pure competition is truly ideal only if all or most industries in the economy are purely competitive and if in addition there is free and easy mobility of productive factors among industries. Otherwise, the relative outputs of different industries will not be such as to maximize consumer satisfaction. There is also some question whether producers in purely competitive industries will generally earn enough to plow back some of their earnings into improved equipment and thus maintain a satisfactory rate of technological progress. Finally, some purely competitive industries have been afflicted with "destructive competition"—the coal industry and the basic agricultural industries, for example. For some historical reason such an industry accumulates excess capacity to the point where sellers suffer chronic losses, and the situation is not corrected by the exit of people and resources from the industry. The invisible hand of the market works too slowly for society to accept. In some cases, notably in agriculture, government has intervened to restrict supply or raise prices. Leaving these qualifications aside, however, the market performance of pure competition furnishes some sort of a standard to which the performance of industries of different structure may be compared.

Monopolistic competition. In the more complex situation of monopolistic competition (atomistic structure with product differentiation) market conduct and performance may be said to follow roughly the tendencies attributed to pure competition. The principal differences are the following. First, individual sellers, because of the differentiation of their products, are able to raise or lower their individual selling prices slightly; they cannot do so by very much, however, because they remain strongly subject to the impersonal forces of the market operating through the general level of prices. Second, rivalry among sellers is likely to involve sales promotion costs as well as the expense of altering products to appeal to buyers. This is

a competitive game that all will play but that nobody, on the average, will win, and the long-run equilibrium price will reflect the added costs involved. In return, however, buyers will get more variety. Third, since not every seller is likely to be equally successful in his sales-promotion and product policies, some will receive profits in excess of a basic interest return on their investment; such profits will come from their success in winning buyers. Monopolistic competition may, like pure competition, include industries that are afflicted with what has been called above destructive competition. This may result not only from a failure to get rid of excess capacity but also from the entry of too many new firms despite the danger of losses.

Monopoly. While single-firm monopolies are rare, except for those subject to public regulation, it is useful to examine the monopolist's market conduct and performance to establish a standard at the other pole from pure competition. As the sole supplier of a distinctive product, the monopolist can set any selling price provided he accepts the sales that correspond to that price. Since the market demand will generally be less the higher the price he sets, the monopolist presumably will set that price that produces the greatest profits, given the relationship of production costs to output. By restricting output he can raise his selling price significantly—an option not open to sellers in atomistic industries.

Monopolistic behaviour

The monopolist will generally charge prices well in excess of production costs and reap profits well above a normal interest return on investment. His output will be substantially smaller, and his price higher, than if he had to meet established market prices as in pure competition. The monopolist may or may not produce at minimal average cost, depending on his cost–output relationship; if he does not, there are no market pressures to force him to do so.

If the monopolist is subject to no threat of entry by a competitor, he will presumably set a selling price that maximizes profits for the industry he monopolizes. If he faces only impeded entry, he may elect to charge a price sufficiently low to discourage entry but above a competitive price—if this will maximize his long-run profits.

Oligopoly. Market conduct and performance in oligopolistic industries generally combine monopolistic and competitive tendencies, with the relative strength of the two tendencies depending roughly on the detailed market structure of the oligopoly.

Rivalry among sellers. In the simplest form of oligopolistic industry, sellers are few and every seller supplies a sufficiently large share of the market so that any feasible and modest change in his policies will appreciably affect the market shares of all his rival sellers and induce them to react or respond. For example, if seller A reduces his selling price below the general level of prices being charged by all sellers sufficiently to permit him to capture significant numbers of customers from his rivals if they hold their selling prices unchanged, they may react by reducing their prices by a similar amount, so that none gains at the expense of others and the group has probably reduced its combined profits. Or seller A's rivals may retaliate by reducing their selling prices more than he did, thus forcing a further reaction from him. Conversely, if seller A increases his selling price above the general level being charged by all sellers (thus tending to lose at least some of his customers to his rivals), they may react by holding their prices unchanged, in which event seller A will probably retract his increase and bring his price back to the previous level. But his rivals may also react by raising their prices as much as seller A raised his, in which case the general level of prices in the industry rises and the combined profits of all sellers are probably increased.

Any seller A in an oligopoly will therefore determine whether or not to alter his price or other market policy in the light of his conjectures about the reactions of his rivals. Correspondingly, his rivals will determine their reactions in the light of their conjectures about what seller A will do in response. The process is not likely to bring the industry price level down to minimal average cost (as in atomistic competition). Many different "equilibrium" levels between the competitive and monopolistic limits are possible, depending on further circumstances.

Thus in an oligopoly viable collusive agreements among rival sellers are quite possible. They may be express agreements established by contract or tacit understandings that develop as a pattern of reactions among sellers to changes in each others' prices or market policies becomes customary. In the United States, express collusive agreements are forbidden by law, but tacit agreements or "gentlemen's understandings" are common in oligopolies. In numerous other Western countries, formal collusive agreements (often called cartels if comprehensive in scope) are legal. Whether tacit or explicit, legal or illegal, one may say that oligopolistic prices tend to be "administered" by sellers, in the senses mentioned above, as distinct from being determined by impersonal market forces.

Sellers' dual aims. The varying market performance of oligopolies results from the fact that individual sellers intrinsically have two conflicting aims. One common desire is to establish among themselves a monopolistic level of price (and of selling costs, etc.), which will maximize their combined profits, giving them the largest "profit pie" to divide. But each seller also has a fundamental antagonism toward rival sellers and wants to maximize his own profits even at the expense of theirs. The relative strengths of these conflicting aims—the maximizing of combined profits and the maximizing of individual profits—will likely depend on how concentrated the oligopoly is, because when sellers are fewer and their individual market shares larger, their rivals' reactions are stronger deterrents to independent actions.

Price behaviour in oligopoly

This is why various sorts of market performance are to be expected in oligopolistic industries. When the entry of other sellers is blockaded, collusive or interdependent behaviour may lead to a full monopoly price. If entry is only impeded, the resulting price may be far enough below the full monopoly level to discourage further entry. But prices are not always what they seem. An announced price that is well above cost may be undercut by clandestine price reductions to individual buyers, bringing the average of actual selling prices down somewhat.

If an oligopolistic industry is made up of a "core" of a few large interdependent sellers plus a "competitive fringe" of several or numerous quite small sellers, the competition of the small sellers may induce the large ones to limit the extent to which they raise their prices.

Price behaviour approaching full monopoly pricing seems to be found mainly in oligopolies having very high seller concentration and blockaded entry. Where these characteristics are less pronounced, prices and profits tend to be lower, though they are likely to be somewhat above the competitive level. A few economists maintain that oligopolistic prices in general do not significantly differ from atomistically competitive prices, but the bulk of statistical evidence does not support them.

In oligopolies in which product differentiation is important, sales-promotion costs and the costs of product improvement or development will display roughly the same variety of tendencies found in pricing. Where there are a few large interdependent sellers, these costs may be restricted to about the same level as those of a single-firm monopolist; on the other hand, rivalry in sales promotion and product development may be sufficient to raise them higher. Oligopolists may also arrive collusively at relatively high uniform selling prices but simultaneously engage in independent nonprice competition (perhaps more so where seller concentration is lower).

WORKABLE COMPETITION

Since the character of market performance varies among industries along with their market characteristics, efforts have been made to devise some practical standard for identifying the sorts of market structure that engender socially satisfactory performance in a given industry. The term workable competition was coined to denote competition that may be considered as leading to a reasonable or socially acceptable approximation to ideal performance in the circumstances of a particular industry. The limits of such an approximation are of course debatable, and so the idea of workable competition must remain elusive because it is basically subjective.

Five attributes of workable competition

Without entering into a complex theoretical discussion of the relationship of individual-industry performance to overall welfare, it is plausible to suggest the following principal attributes of workable performance in an industry: (1) In the long term, selling price on the average should be equal to or not significantly above average costs of production, so that profits do not appreciably exceed a normal interest return on investment. Prices should be responsive to basic reductions in costs. (2) In so far as average costs of production are affected by the scales or capacities of plants and firms, the preponderance of industry output should be from plants and firms of the most efficient scale or with closely comparable technical efficiency. (3) The industry should not have chronic excess capacity—*i.e.*, significant plant capacity that is persistently unused even in periods of high general economic activity. (4) The industry's sales-promotion costs should not be substantially greater than needed to keep buyers informed of the availability, characteristics, and prices of products. (5) The industry should be adequately progressive in introducing more economical production techniques and improved products—balancing the costs of progress with the gains.

While the first three of these attributes are easier to appraise than the others, certain generalizations are possible concerning the workability of different market structures: (1) Unregulated single-firm monopolies tend to generate unworkable market performance, mainly in the form of output restriction, prices well above costs, and consequent excess profits. They have undesirable effects on the uses to which resources are put and on income distribution. (2) Oligopolies with high seller concentration and also very high barriers to entry tend toward unworkable performance, like that of single-firm monopoly. In general, however, they do not show significant degrees of technical inefficiency resulting from inefficient plant scales or excess capacity. (3) Oligopolies with fairly high seller concentration but only moderate barriers to entry are also prone to unworkable performance of the sort just mentioned, but not to as high a degree. (4) Oligopolies with only moderate seller concentration and moderate-to-low barriers to entry tend toward workable performance both in price–cost relations and in technical efficiency, except that some of them may have recurrent chronic excess capacity due to periodic overentry. (If cartels are legalized and their provisions are not rigorously controlled by government, the last two categories of oligopoly may have the same sort of unworkable performance as do very highly concentrated oligopolies.) (5) Industries of atomistic structure tend generally toward workable performance unless they suffer from destructive competition as described above.

Product differentiation and promotion

In industries with significant differentiation of products among sellers—and especially in oligopolies of this sort—there is a tendency for minor but significant fractions of income to be devoted to persuasive (as distinct from informational) advertising and other sales promotion and also to more or less idle variations of product design, with the result that resources are in a sense "wasted" and costs increased.

By the criteria of workable competition, a purely rational society would presumably favour industries with moderate to low seller concentration, moderate to low barriers to entry, and without extreme product differentiation—all this from the standpoint of enhancing overall material welfare. The argument that oligopolistic and atomistic industries generally need legal protection from destructive competition may be discarded on the basis of evidence. Price and other market warfare in such industries has been extremely rare in industrial countries in the last 50 years.

(J.S.B./Ed.)

Production: the output of the factors of production

In economics, the theory of production is an effort to explain the principles by which a business firm decides how much of each commodity that it sells (its "outputs" or "products") it will produce, and how much of each kind of labour, raw material, fixed capital good, etc., that it employs (its "inputs" or "factors of production") it will use. The theory involves some of the most fundamental principles of economics. These include the relationship between the prices of commodities and the prices (or wages or rents) of the productive factors used to produce them and also the relationships between the prices of commodities and productive factors, on the one hand, and the quantities of these commodities and productive factors that are produced or used, on the other.

The various decisions a business enterprise makes about its productive activities can be classified into three layers of increasing complexity. The first layer includes decisions about methods of producing a given quantity of the output in a plant of given size and equipment. It involves the problem of what is called short-run cost minimization. The second layer, including the determination of the most profitable quantities of products to produce in any given plant, deals with what is called short-run profit maximization. The third layer, concerning the determination of the most profitable size and equipment of plant, relates to what is called long-run profit maximization.

MINIMIZATION OF SHORT-RUN COSTS

The relationship between input and output

The production function. However much of a commodity a business firm produces, it endeavours to produce it as cheaply as possible. Taking the quality of the product and the prices of the productive factors as given, which is the usual situation, the firm's task is to determine the cheapest combination of factors of production that can produce the desired output. This task is best understood in terms of what is called the production function, *i.e.*, an equation that expresses the relationship between the quantities of factors employed and the amount of product obtained. It states the amount of product that can be obtained from each and every combination of factors. This relationship can be written mathematically as $y = f(x_1, x_2, \ldots, x_n; k_1, k_2, \ldots, k_m)$. Here, y denotes the quantity of output. The firm is presumed to use n variable factors of production; that is, factors like hourly paid production workers and raw materials, the quantities of which can be increased or decreased. In the formula the quantity of the first variable factor is denoted by x_1 and so on. The firm is also presumed to use m fixed factors, or factors like fixed machinery, salaried staff, etc., the quantities of which cannot be varied readily or habitually. The available quantity of the first fixed factor is indicated in the formal by k_1 and so on. The entire formula expresses the amount of output that results when specified quantities of factors are employed. It must be noted that though the quantities of the factors determine the quantity of output, the reverse is not true, and as a general rule there will be many combinations of productive factors that could be used to produce the same output. Finding the cheapest of these is the problem of cost minimization.

The cost of production is simply the sum of the costs of all of the various factors. It can be written:

$$C = p_1x_1 + \cdots + p_nx_n + r_1k_1 + \cdots + r_nk_n,$$

in which p_1 denotes the price of a unit of the first variable factor, r_1 denotes the annual cost of owning and maintaining the first fixed factor, and so on. Here again one group of terms, the first, covers variable cost (roughly "direct costs" in accounting terminology), which can be changed readily; another group, the second, covers fixed cost (accountants' "overhead costs"), which includes items not easily varied. The discussion will deal first with variable cost.

The principles involved in selecting the cheapest combination of variable factors can be seen in terms of a simple example. If a firm manufactures gold necklace chains in such a way that there are only two variable factors, labour (specifically, goldsmith-hours) and gold wire, the production function for such a firm will be $y = f(x_1, x_2; k)$, in which the symbol k is included simply as a reminder that the number of chains producible by x_1 feet of gold wire and x_2 goldsmith-hours depends on the amount of machinery and other fixed capital available. Since there are only two variable factors, this production function can be portrayed graphically in a figure known as an isoquant diagram (Figure 8). In the graph, goldsmith-hours per

month are plotted horizontally and the number of feet of gold wire used per month vertically. Each of the curved lines, called an isoquant, will then represent a certain number of necklace chains produced. The data displayed show that 100 goldsmith-hours plus 900 feet of gold wire can produce 200 necklace chains. But there are other combinations of variable inputs that could also produce 200 necklace chains per month. If the goldsmiths work more carefully and slowly, they can produce 200 chains from 850 feet of wire; but to produce so many chains more goldsmith-hours will be required, perhaps 130. The isoquant labelled "200" shows all the combinations of the variable inputs that will just suffice to produce 200 chains. The other two isoquants shown are interpreted similarly. It is obvious that many more isoquants, in principle an infinite number, could also be drawn. This diagram is a graphic display of the relationships expressed in the production function.

Substitution of factors. The isoquants also illustrate an important economic phenomenon: that of factor substitution. This means that one variable factor can be substituted for others; as a general rule a more lavish use of one variable factor will permit an unchanged amount of output to be produced with fewer units of some or all of the others. In the example above, labour was literally as good as gold and could be substituted for it. If it were not for factor substitution there would be no room for further decision after y, the number of chains to be produced, had been established.

The shape of the isoquants shown, for which there is a good deal of empirical support, is very important. In moving along any one isoquant, the more of one factor that is employed, the less of the other will be needed to maintain the stated output; this is the graphic representation of factor substitutability. But there is a corollary: the more of one factor that is employed, the less it will be possible to reduce the use of the other by using more of the first. This is the property known as "diminishing marginal rates of substitution." The marginal rate of substitution of factor 1 for factor 2 is the number of units by which x_1 can be reduced per unit increase in x_2 output remaining unchanged. In the diagram, if feet of gold wire

Diminishing marginal rate of substitution

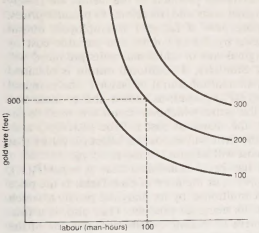


Figure 8: Isoquant diagram of hours of labour and feet of gold wire used per month.

are indicated by x_1 and goldsmith-hours by x_2, then the marginal rate of substitution is shown by the steepness (the negative of the slope) of the isoquant; and it will be seen that it diminishes steadily as x_2 increases because it becomes harder and harder to economize on the use of gold simply by taking more care. The remainder of the analysis rests heavily on the assumption that diminishing marginal rates of substitution are characteristic of the production process generally.

The cost data and the technological data can now be brought together. The variable cost of using x_1, x_2 units of the factors of production is written $p_1x_1 + p_2x_2$, and this information can be added to the isoquant diagram (Figure 9). The straight line labelled v_2, called the v-isocost line, shows all the combinations of input that can be purchased for a specified variable cost, v_2. The other two isocost lines shown are interpreted similarly. The general formula for an isocost line is $p_1x_1 + p_2x_2 = v$, in which v is some particular variable cost. The slope of an isocost line is found by dividing p_2 by p_1 and depends only on the ratio of the prices of the two factors.

Three isocost lines are shown, corresponding to variable costs amounting to v_1, v_2, and v_3. If 200 units are to be produced, expenditure of v_1 on variable factors will not suffice since the v_1-isocost line never reaches the isoquant for 200 units. An expenditure of v_3 is more than sufficient; and v_2 is the lowest variable cost for which 200 units can be produced. Thus v_2 is found to be the minimum variable cost of producing 200 units (as v_3 is of 300 units) and the coordinates of the point where the v_2 isocost line touches the 200-unit isoquant are the quantities of the two factors

The minimization of costs

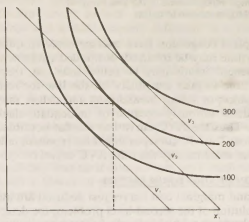


Figure 9: Isoquant diagram for two factors of production, x_1 and x_2 (see text).

that will be used when 200 units are to be produced and the prices of the two factors are in the ratio p_2/p_1. It may be noted that the cheapest combination for the production of any quantity will be found at the point at which the relevant isoquant is tangent to an isocost line. Thus, since the slope of an isoquant is given by the marginal rate of substitution, any firm trying to produce as cheaply as possible will always purchase or hire factors in quantities such that the marginal rate of substitution will equal the ratio of their prices.

The isoquant–isocost diagram (or the corresponding solution by the alternative means of the calculus) solves the short-run cost minimization problem by determining the least-cost combination of variable factors that can produce a given output in a given plant. The variable cost incurred when the least-cost combination of inputs is used in conjunction with a given outfit of fixed equipment is called the variable cost of that quantity of output and denoted $VC(y)$. The total cost incurred, variable plus fixed, is the short-run cost of that output, denoted $SRC(y)$. Clearly $SRC(y) = VC(y) + R(K)$, in which the second term symbolizes the sum of the annual costs of the fixed factors available.

Marginal cost. Two other concepts now become important. The average variable cost, written $AVC(y)$, is the variable cost per unit of output. Algebraically, $AVC(y) = VC(y)/y$. The marginal variable cost, or simply marginal cost $[MC(y)]$ is, roughly, the increase in variable cost incurred when output is increased by one unit; *i.e.*, $MC(y) = VC(y+1) - VC(y)$. Though for theoretical purposes a more precise definition can be obtained by regarding $VC(y)$ as a continuous function of output, this is not necessary in the present case.

The usual behaviour of average and marginal variable costs in response to changes in the level of output from a given fixed plant is shown in Figure 10. In this figure costs (in dollars per unit) are measured vertically and output (in units per year) is shown horizontally. The figure is drawn for some particular fixed plant, and it can be seen that average costs are fairly high for very low levels of output relative to the size of the plant, largely because there is not enough work to keep a well-balanced work force fully occupied. People are either idle much of the time or shifting expensively, from job to job. As output increases from a low level, average costs decline to a low plateau. But as the capacity of the plant is approached, the inefficiencies

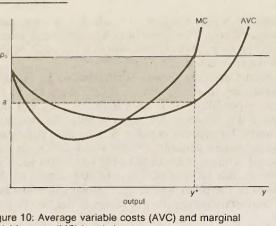


Figure 10: Average variable costs (AVC) and marginal variable costs (MC) in relation to output.

incident on plant congestion force average costs up quite rapidly. Overtime may be incurred, outmoded equipment and inexperienced hands may be called into use, there may not be time to take machinery off the line for routine maintenance; or minor breakdowns and delays may disrupt schedules seriously because of inadequate slack and reserves. Thus the AVC curve has the flat-bottomed U-shape shown. The MC curve, as might be expected, falls faster and rises more rapidly than the AVC curve.

MAXIMIZATION OF SHORT-RUN PROFITS

How the firm maximizes its profit

The average and marginal cost curves just deduced are the keys to the solution of the second-level problem, the determination of the most profitable level of output to produce in a given plant. The only additional datum needed is the price of the product, say p_0.

The most profitable amount of output may be found by using these data. If the marginal cost of any given output (y) is less than the price, sales revenues will increase more than costs if output is increased by one unit (or even a few more); and profits will rise. Contrariwise, if the marginal cost is greater than the price, profits will be increased by cutting back output by at least one unit. It then follows that the output that maximizes profits is the one for which $MC(y) = p_0$. This is the second basic finding: in response to any price the profit-maximizing firm will produce and offer the quantity for which the marginal cost equals that price.

Such a conclusion is shown in Figure 10. In response to the price, p_0, shown, the firm will offer the quantity y^* given by the value of y for which the ordinate of the MC curve equals the price. If a denotes the corresponding average variable cost, net revenue per unit will be equal to $p_0 - a$, and the total excess of revenues over variable costs will be $y^*(p_0 - a)$, which is represented graphically by the shaded rectangle in the figure.

Marginal cost and price. The conclusion that marginal cost tends to equal price is important in that it shows how the quantity of output produced by a firm is influenced by the market price. If the market price is lower than the lowest point on the average variable cost curve, the firm will "cut its losses" by not producing anything. At any higher market price, the firm will produce the quantity for which marginal cost equals that price. Thus the quantity that the firm will produce in response to any price can be found in Figure 10 by reading the marginal cost curve, and for this reason the marginal cost curve is said to be the short-run supply curve for the firm.

The short-run supply curve for a product—that is, the total amount that all the firms producing it will produce in response to any market price—follows immediately, and is seen to be the sum of the short-run supply curves (or marginal cost curves, except when the price is below the bottoms of the average variable cost curves for some firms) of all the firms in the industry. This curve is of fundamental importance for economic analysis, for together with the demand curve for the product it determines the market price of the commodity and the amount that will be produced and purchased.

One pitfall must, however, be noted. In the demonstration of the supply curves for the firms, and hence of the industry, it was assumed that factor prices were fixed. Though this is fair enough for a single firm, the fact is that if all firms together attempt to increase their outputs in response to an increase in the price of the product, they are likely to bid up the prices of some or all of the factors of production that they use. In that event the product supply curve as calculated will overstate the increase in output that will be elicited by an increase in price. A more sophisticated type of supply curve, incorporating induced changes in factor prices, is therefore necessary. Such curves are discussed in the standard literature of this subject.

Marginal product. It is now possible to derive the relationship between product prices and factor prices, which is the basis of the theory of income distribution. To this end, the marginal product of a factor is defined as the amount that output would be increased if one more unit of the factor were employed, all other circumstances remaining the same. Algebraically, it may be expressed as the difference between the product of a given amount of the factor and the product when that factor is increased by an additional unit. Thus if $MP_1(x_1)$ denotes the marginal product of factor 1 when x_1 units are employed, then $MP_1(x_1) = f(x_1 + 1, x_2, \ldots, x_n; k) - f(x_1, x_2, \ldots, x_n; k)$. The marginal products are closely related to the marginal rates of substitution previously defined. If an additional unit of factor 1 will increase output by f_1 units, for example, then one more unit of output can be obtained by employing $1/f_1$ more units of factor 1. Similarly, if the marginal product of factor 2 is f_2, then output will fall by one unit if the use of factor 2 is reduced by $1/f_2$ units. Thus output will remain unchanged, to a good approximation, if $1/f_1$ units of factor 1 are used to replace $1/f_2$ units of factor 2. The marginal rate of substitution is therefore f_2/f_1, or the ratio of the marginal products of the two factors. It has already been shown that the marginal rate of substitution also equals the ratio of the prices of the factors, and it therefore follows that the prices (or wages) of the factors are proportional to their marginal products.

The returns to the factors of production

This is one of the most significant theoretical findings in economics. To restate it briefly: factors of production are paid in proportion to their marginal products. This is not a question of social equity but merely a consequence of the efforts of businessmen to produce as cheaply as possible.

Further, the marginal products of the factors are closely related to marginal costs and, therefore, to product prices. For if one more unit of factor 1 is employed, output will be increased by $MP_1(x_1)$ units and variable cost by p_1; so the marginal cost of additional units produced will be $p_1/MP_1(x_1)$. Similarly, if additional output is obtained by employing an additional unit of factor 2, the marginal cost will be $p_2/MP_2(x_2)$. But, as shown above, these two numbers are the same; whichever factor i is used to increase output, the marginal cost will be $p_i/MP_i(x_i)$ and, furthermore, the firm will choose its output level so that the marginal cost will be equal to the price, p_0.

Therefore it has been established that $p_1 = p_0 MP_1(x_1)$, $p_2 = p_0 MP_2(x_2)$, . . . , or the price of each factor is the price of the product multiplied by its marginal product, which is the value of its marginal product. This, also, is a fundamental theorem of income distribution and one of the most significant theorems in economics. Its logic can be perceived directly. If the equality is violated for any factor, the businessman can increase his profits either by hiring units of the factor or by laying them off until the equality is satisfied, and presumably the businessman will do so.

The theory of production decisions in the short run, as just outlined, leads to two conclusions (of fundamental importance throughout the field of economics) about the responses of business firms to the market prices of the commodities they produce and the factors of production they buy or hire: (1) the firm will produce the quantity of its product for which the marginal cost is equal to the market price and (2) it will purchase or hire factors of production in such quantities that the price of the commodity produced multiplied by the marginal product of the factor will be equal to the cost of a unit of the factor. The first explains the supply curves of the commodities produced in an economy. Though the conclusions were deduced within the context of a firm that uses two factors of production, they are clearly applicable in general.

MAXIMIZATION OF LONG-RUN PROFITS

The relationship between the short run and the long run

The theory of long-run profit-maximizing behaviour rests on the short-run theory that has just been presented but is considerably more complex because of two features: (1) long-run cost curves, to be defined below, are more varied in shape than the corresponding short-run cost curves, and (2) the long-run behaviour of an industry cannot be deduced simply from the long-run behaviour of the firms in it because the roster of firms is subject to change. It is of the essence of long-run adjustments that they take place by the addition or dismantling of fixed productive capacity by both established firms and new or recently created firms.

At any one time an established firm with an existing plant will make its short-run decisions by comparing the ruling price of its commodity with cost curves corresponding to that plant. If the price is so high that the firm is operating on the rising leg of its short-run cost curve, its marginal costs will be high—higher than its average costs—and it will be enjoying operating profits, as shown in Figure 10. The firm will then consider whether it could increase its profits by enlarging its plant. The effect of plant enlargement is to reduce the variable cost of producing high levels of output by reducing the strain on limited production facilities, at the expense of increasing the level of fixed costs.

In response to any level of output that it expects to continue for some time, the firm will desire and eventually acquire the fixed plant for which the short-run costs of that level of output are as low as possible. This leads to the concept of the long-run cost curve: the long-run costs of any level of output are the short-run costs of producing that output in the plant that makes those short-run costs as low as possible. These result from balancing the fixed costs entailed by any plant against the short-run costs of producing in that plant. The long-run costs of producing y are denoted by $LRC(y)$. The average long-run cost of y is the long-run cost per unit of y [algebraically $LAC(y) = LRC(y)/y$]. The marginal long-run cost is the increase in long-run cost resulting from an increase of one unit in the level of output. It represents a combination of short-run and long-run adjustments to a slight increase in the rate of output. It can be shown that the long-run marginal cost equals the marginal cost as previously defined when the cost-minimizing fixed plant is used.

Classes of long-run cost curves

Cost curves appropriate for long-run analysis are more varied in shape than short-run cost curves and fall into three broad classes. In constant-cost industries, average cost is about the same at all levels of output except the very lowest. Constant costs prevail in manufacturing industries in which capacity is expanded by replicating facilities without changing the technique of production, as a cotton mill expands by increasing the number of spindles. In decreasing-cost industries, average cost declines as the rate of output grows, at least until the plant is large enough to supply an appreciable fraction of its market. Decreasing costs are characteristic of manufacturing in which heavy, automated machinery is economical for large volumes of output. Automobile and steel manufacturing are leading examples. Decreasing costs are inconsistent with competitive conditions, since they permit a few large firms to drive all smaller competitors out of business. Finally, in increasing-cost industries average costs rise with the volume of output generally because the firm cannot obtain additional fixed capacity that is as efficient as the plant it already has. The most important examples are agriculture and extractive industries.

CRITICISMS OF THE THEORY

The theory of production has been subject to much criticism. One objection is that the concept of the production function is not derived from observation or practice. Even the most sophisticated firms do not know the direct functional relationship between their basic raw inputs and their ultimate outputs. This objection can be got around by applying the recently developed techniques of linear programming, which employ observable data without recourse to the production function and lead to practically the same conclusions.

On another level the theory has been charged with excessive simplification. It assumes that there are no changes in the rest of the economy while individual firms and industries are making the adjustments described in the theory; it neglects changes in the technique of production; and it pays no attention to the risks and uncertainties that becloud all business decisions. These criticisms are especially damaging to the theory of long-run profit maximization. On still another level, critics of the theory maintain that businessmen are not always concerned with maximizing profits or minimizing costs.

Though all of the criticisms have merit, the simplified theory of production does nevertheless indicate some basic forces and tendencies operating in the economy. The theorems should be understood as conditions that the economy tends toward, rather than conditions that are always and instantaneously achieved. It is rare for them to be attained exactly, but it is just as rare for substantial violations of the theorems to endure.

Only the simplest aspects of the theory were described above. Without much difficulty it could be extended to cover firms that produce more than one product, as almost all firms do. With more difficulty it could be applied to firms whose decisions affect the prices at which they sell and buy (monopoly, monopolistic competition, monopsony). The behaviour of other firms that recognize the possibility that their competitors may retaliate (oligopoly) is still a theory of production subject to controversy and research. (R.D.)

Distribution: the shares of the factors of production

The factors of production, as suggested earlier, are the economic resources, both human and other, which, if properly utilized, will bring about a flow or output of goods and services. The factors are commonly classified into three groups: capital, labour, and land. The first, in the simplest sense, refers to all the "produced" instruments of production—the factories, their equipment, their stocks of raw materials and finished goods, houses, trade facilities, and so on; the owners of capital receive their income in various possible forms, profits and interest being the usual ones. The factor of labour represents all those productive resources that can be applied only at the cost of human effort; the wage (or salary) is the form of payment for use of this factor. The factor of land represents resources whose supply is low in relation to demand and cannot be increased as the result of production; the income derived from the ownership of this factor is known as economic rent.

Capital, labour, and land

Distribution, in economics, generally refers to (1) explanations of how prices for the services of the different factors of production are determined; (2) explanations of how the total product of the economy is divided among the various factors, and (3) descriptions of the ways in which the income is divided among various income classes or groups of persons.

CAPITAL AND INTEREST

Capital in economics is a word of many meanings. They all imply that capital is a "stock" by contrast with income, which is a "flow." In its broadest possible sense, capital includes the human population; nonmaterial elements such as skills, abilities, and education; land, buildings, machines, equipment of all kinds; and all stocks of goods—finished or unfinished—in the hands of both firms and households.

In the business world the word capital usually refers to an item in the balance sheet representing that part of the net worth of an enterprise that has not been produced through the operations of the enterprise. In economics the word capital is generally confined to "real" as opposed to merely "financial" assets. Different as the two concepts may seem, they are not unrelated. If all balance sheets were consolidated in a closed economic system, all debts would be cancelled out because every debt is an asset in one balance sheet and a liability in another. What is left in the consolidated balance sheet, therefore, is a value of all the

real assets of a society on one side and its total net worth on the other. This is the economist's concept of capital.

Kinds of capital

A distinction may be made between goods in the hands of firms and goods in the hands of households, and attempts have been made to confine the term capital structure to the former. There is also a distinction between goods that have been produced and goods that are gifts of nature; attempts have been made to confine the term capital to the former, though the distinction is hard to maintain in practice. Another important distinction is between the stock of human beings (and their abilities) and the stock of nonhuman elements. In a slave society human beings are counted as capital in the same way as livestock or machines. In a free society each man is his own slave—the value of his body and mind is not, therefore, an article of commerce and does not get into the accounting system. In strict logic persons should continue to be regarded as part of the capital of a society; but in practice the distinction between the part of the total stock that enters into the accounting system, and the part that does not, is so important that it is not surprising that many writers have excluded persons from the capital stock.

Another distinction that has some historical importance is that between circulating and fixed capital. Fixed capital is usually defined as that which does not change its form in the course of the process of production, such as land, buildings, and machines. Circulating capital consists of goods in process, raw materials, and stocks of finished goods waiting to be sold; these goods must either be transformed, as when wheat is ground into flour, or they must change ownership, as when a stock of goods is sold. This distinction, like many others, is not always easy to maintain. Nevertheless, it represents a rough approach to an important problem of the relative structure of capital; that is, of the proportions in which goods of various kinds are found. The stock of real capital exhibits strong complementarities. A machine is of no use without a skilled operator and without raw materials for it to work on.

The classical theory of capital. Although ancient and medieval writers were interested in the ethics of interest and usury, the concept of capital as such did not rise to prominence in economic thought before the classical economists (Adam Smith, David Ricardo, Nassau Senior, and John Stuart Mill).

The ideas of economists about capital

Adam Smith laid great stress on the role played by the accumulation of a stock of capital in facilitating the division of labour economics and in increasing the productivity of labour in general. He recognized clearly that accumulation proceeds from an excess of production over consumption. He distinguished between productive labour, which creates objects of capital, and unproductive labour (services), the fruits of which are enjoyed immediately. His thought was strongly coloured by observation of the annual agricultural cycle. The end of the harvest saw society with a given stock of grain. This stock was in the possession of the capitalists. A certain portion of it they reserved for their own consumption and for the consumption of their menial servants, the rest was used to feed "productive labourers" during the ensuing year. As a result, by the end of the next harvest the barns were full again and the stock had replaced itself, perhaps with something left over. The stock that the capitalists did not reserve for their own use was the "wages fund"—the more grain there was in the barn in October the sharper the competition of capitalists for workers, and the higher real wages would be in the year to come. The picture is a crude one, of course, and does not indicate the complexity of the relationship between stocks and flows in an industrial society. The last of the classical economists, John Stuart Mill, was forced to abandon the wages-fund theory. Nevertheless, the wages fund is a crude representation of some real but complex relationships, and the theory reappears in a more sophisticated form in later writers.

The classical economists distinguished three categories of income—wages, profit, and rent—and identified these with three factors of production—labour, capital, and land. David Ricardo especially made a sharp distinction between capital as "produced means of production," and land as the "original and indestructible powers of the soil." In modern economics this distinction has become blurred.

The Austrian school. About 1870 a new school developed, sometimes called the Austrian school from the fact that many of its principal members taught in Vienna, but perhaps better called the Marginalist school. The movement itself was thoroughly international, and included such figures as William Stanley Jevons in England and Léon Walras in France. The so-called Austrian theory of capital is mainly based on the work of Eugen Böhm-Bawerk. His *Positive Theory of Capital* (1889) set off a controversy that has not yet subsided. In the Austrian view the economic process consisted of the embodiment of "original factors of production" in capital goods of greater or lesser length of life that then yielded value or utility as they were consumed. Between the original embodiment of the factor and the final fruition in consumption lay an interval of time known as the period of production. In an equilibrium population it can easily be shown that the total population (capital stock) equals the annual number of births or deaths (income) multiplied by the average length of life (period of production). The longer the period of production, therefore, the more capital goods there will be per unit of income. If the period of production is constant, income depends directly on the amount of capital previously accumulated. Here is the wages fund in a new form. Unfortunately, the usefulness of Böhm-Bawerk's theory is much impaired by the fact that it is confined to equilibrium states. The great problems of capital theory are dynamic in character, and comparative statics throws only a dim light on them.

Marginalist and Keynesian theories. The Marginalist school culminated in the work of three men—P.H. Wicksteed in England, Knut Wicksell in Sweden, and Irving Fisher in the United States. The last two especially gave the Austrian theory clear mathematical expression. Perhaps the greatest contribution of the Austrian theory was its recognition of the importance of the valuation problem in the relation of capital to interest. From the mere fact that physical capital produces an income stream, there is no explanation of the phenomenon of interest, for the question is why the value of a piece of physical capital should be less than the total of future values that are expected to accrue from it. The theory also makes a contribution to the problem of rational choice in situations involving waiting or maturing. The best example is that of slowly maturing goods such as wines or timber. There is a problem here of the best time to draw wine or to cut down a tree. According to the marginal theory this is at the time when the rate of net value growth of the item is just equal to the rate of interest, or the rate of return in alternative investments. Thus, if a tree or a wine is increasing in value at the rate of 7 percent per annum when the rate of interest is 6 percent it still pays to be patient and let it grow or mature. The longer it grows, however, the less the rate of value growth, and when the rate of value growth has fallen to the rate of interest, then is the time to reap the fruits of patience.

Growth and marginal value

The contributions of John Maynard (Lord) Keynes to capital theory are incidental rather than fundamental. Nevertheless, the "Keynesian revolution" had an impact on this area of economic thought as on most others. It overthrew the traditional assumption of most economists that savings were automatically invested. The great contribution of Keynes, then, is the recognition that the attempt to save does not automatically result in the accumulation of capital. A decision to restrict consumption is only a decision to accumulate capital if the volume of production is constant. If abstention from consumption itself results in a diminution of production, then accumulation (production minus consumption) is correspondingly reduced.

Important theoretical questions

Later thinking. The theory of capital was not a matter of primary concern to economists in the late 20th century, though some revival of interest occurred in the late 1950s. Nevertheless, certain problems remain of perennial interest. They may be grouped as follows.

Heterogeneous goods. First are the problems involved in measuring aggregates of goods. Real capital includes everything from screwdrivers to continuous strip-rolling mills. A single measure of total real capital can be achieved only

if each item can be expressed in a common denominator such as a given monetary unit (*e.g.*, dollars, sterling, francs, pesos, etc.). The problem becomes particularly complicated in periods of rapid technical change when there is change not only in the relative values of products but in the nature of the list itself. Only approximate solutions can be found to this problem, and no completely satisfactory measure is ever possible.

A related problem that has aroused considerable interest among accountants is how to value capital assets that have no fixed price. In the conventional balance sheet the value of some items is based on their cost at an earlier period than that of others. When the general level of prices is changing this means that different items are valued in monetary units of different purchasing power. The problem is particularly acute in the valuation of inventory. Under the more conventional "FIFO" (First In, First Out) system, inventory is valued at the cost (purchase price) of the latest purchases. This leads to an inflation of inventory values, and therefore of accounting profits, in time of rising prices (and a corresponding deflation under falling prices), which may be an exaggeration of the long-run position of the firm. This may be partially avoided by a competing system of valuation known as LIFO (Last In, First Out), in which inventory is valued at the purchase price of the earliest purchases. This avoids the fluctuations caused by short-run price-level changes, but it fails to record changes in real long-run values. There seems to be no completely satisfactory solution to this problem, and it is wise to recognize the fact that any single figure of capital value that purports to represent a complex, many-dimensional reality will need careful interpretation.

The accumulation process. A second problem concerns the factors that determine the rate of accumulation of capital; that is, the rate of investment. It has been seen that investment in real terms is the difference between production and consumption. The classical economist laid great stress on frugality as the principal source of capital accumulation. If production is constant it is true that the only way to increase accumulation is by the reduction of consumption. Keynes shifted the emphasis from the reduction of consumption to the increase of production, and regarded the decision to produce investment goods as the principal factor in determining the rate of growth of capital. In modern theories of economic development great stress is laid on the problem of the structure of production—the relative proportions of different kinds of activity. The advocates of "balanced growth" emphasize the need for a developing country to invest in a wide range of related and cooperative enterprises, public as well as private. There is no point in building factories and machines, they say, if the educational system does not provide a labour force capable of using them. There is also, however, a case to be made for "unbalanced growth," in the sense that growth in one part of the economy frequently stimulates growth in other parts. A big investment in mining or in hydroelectric power, for example, creates strains on the whole society, which result in growth responses in the complementary sectors. The relation of inflation to economic growth and investment is an important though difficult problem. There seems to be little doubt that deflation, mainly because it shifts the distribution of income away from the profit maker toward the *rentier* and bondholder, has a deleterious effect on investment and the growth of capital. In 1932, for instance, real investment had practically ceased in the United States. It is less clear at what point inflation becomes harmful to investment. In countries where there has been long continuing inflation there seems to be some evidence that the structure of investment is distorted. Too much goes into apartment houses and factories and not enough into schools and communications.

Effects of deflation and inflation

Capital and time. A third problem that exists in capital theory is that of the period of production and the time structure of the economic process. This cannot be solved by the simple formulas of the Austrian school. Nevertheless, the problem is a real one and there is still a need for more useful theoretical formulations of it. Decisions taken today have results extending far into the future. Similarly, the data of today's decisions are the result of decisions that were taken long in the past. The existing capital structure is the embodiment of past decisions and the raw material of present decisions. The incompatibility of decisions is frequently not discovered at the time they are made because of the lapse of time between the decision and its consequences. It is tempting to regard the cyclical structure of human history, whether the business cycle or the war cycle, as a process by which the consequences of bad decisions accumulate until some kind of crisis point is reached. The crisis (a war or a depression) redistributes power in the society and so leads to a new period of accumulating, but hidden, stress. In this process, distortion in the capital structure is of great importance.

Capital and income. A fourth problem to be considered is the relationship that exists between the stocks and the flows of a society, or in a narrower sense the relation between capital and income. Income, like capital, is a concept that is capable of many definitions; a useful approach to the concept of income is to regard it as the gross addition to capital in a given period. For any economic unit, whether a firm or an individual, income may be measured by that hypothetical amount of consumption that would leave capital intact. In real terms this is practically identical with the concept of production. The total flow of income is closely related to both the quantity and the structure of capital; the total real income of a society depends on the size and the skills of its population, and on the nature and the extent of the equipment with which they have to work. The most important single measure of economic well-being is real income per person; this is closely related to the productivity of labour, and this in turn is closely related to capital per person, especially if the results of investment in human resources, skills, and education are included in the capital stock.

The nature of interest

Interest. Historically, the concept of capital has been so closely bound to the concept of interest that it seems wise to take these two topics together, even though in the modern view it is capital and income rather than capital and interest that are the related concepts.

Interest as a form of income may be defined as income that is received as a result of the possession of contractual obligations for payment on the part of another. Interest, in other words, is income that is received as a result of the ownership of a bond, a promissory note, or some other instrument that represents a promise on the part of some other party to pay sums in the future. The obligations may take many forms. In the case of the perpetuity, the undertaking is to pay a certain sum each year or other interval of time for the indefinite future. A bond with a date of maturity usually involves a promise to pay a certain sum each year for a given number of years, and then a larger sum on the terminal date. A promissory note frequently consists of a promise to pay a single sum at a date that is some time in the future.

If $a_1, a_2, \ldots a_n$ are the sums received by the bondholder in years $1, 2 \ldots n$, and if P_0 is the present value in year 0, or the sum for which the bond is purchased, the rate of interest r in the whole transaction is given by the equation

$$P_0 = a_1(1+r)^{-1} + a_2(1+r)^{-2} + \cdots + a_n(1+r)^{-n}.$$

There is no general solution for this equation, though in practice it can be solved easily by successive approximation, and in special cases the equation reduces to much simpler forms. In the case of a promissory note, for instance, the equation reduces to the form

$$P_0 = a_n(1+r)^{-n}, \text{ or } \log(1+r) = \frac{\log a_n - \log P_0}{n}$$

where a_n is the single promised payment. In the case of a perpetuity with an annual payment of a, the formula reduces to

$$P_0 = a[(1+r)^{-1} + (1+r)^{-2} + \cdots \text{ to inf.}] = \frac{a}{r}$$

whence $r = \dfrac{a}{P_0}$.

Thus if one had to pay $200 to purchase a perpetual annuity of $5 per annum, the rate of interest would be $2\frac{1}{2}$ percent.

It should be observed that the dimensions of the rate of

interest are those of a rate of growth. The rate of interest is not a price or ratio of exchange; it is not itself determined in the market. What is determined in the market is the price of contractual obligations or "bonds." The higher the price of a given contractual obligation, the lower the rate of interest on it. Suppose, for instance, that one has a promissory note that is a promise to pay one $100 in one year's time. If I buy this for $100 now, the rate of interest is zero; if I buy it for $95 now the rate of interest is a little over 5 percent; if I buy it for $90 now, the rate of interest is about 11 percent. The rate of interest may be defined as the gross rate of growth of capital in a contractual obligation.

A distinction is usually made between interest and profit as forms of income. In ordinary speech, profit usually refers to income derived from the ownership of aggregates or assets of all kinds organized in an enterprise. This aggregate is described by a balance sheet. In the course of the operations of the enterprise, the net worth grows, and profit is the gross growth of net worth. Stocks, as opposed to bonds, usually imply a claim on the profits of some enterprise.

Interest theory since Aristotle

The development of interest theory. In ancient and medieval times the main focus of inquiry into the theory of interest was ethical, and the principal question was the moral justification of interest. On the whole, the taking of interest was regarded unfavourably by both classical and medieval writers. Aristotle regarded money as "barren" and the medieval schoolmen were hostile to usury. Nevertheless, where interest fulfilled a useful social function elaborate rationalizations were developed for it. Among the classical economists, the focus of attention shifted away from ethical justification toward the problem of mechanical equilibrium. The question then became this: Is there any equilibrium rate of interest or rate of profit in the sense that where actual rates are above or below this, forces are brought into play, tending to change them toward the equilibrium? The classical economists did not provide any clear solution for this problem. They believed that the rate of interest simply followed the rate of profit, for people would not borrow or incur contractual obligations unless they could earn something more than the cost of the borrowing by investing the proceeds in enterprises or aggregates of real capital. They believed that the growth of capital itself would tend to reduce the rate of profit because of the competition of the capitalists. This doctrine is important in the Marxian dynamics in which the struggle of capital to avoid a falling rate of profit is seen as a critical factor leading, for instance, to unemployment, foreign investment, and imperialism.

In the framework of classical economics, the work of Nassau Senior deserves mention. He raised the question whether profit or interest were "paid for" anything; that is, whether there was any identifiable contribution to the general product of society that would not be forthcoming if this form of income were not paid. He identified such a function and called it abstinence. Karl Marx denied the existence of any such function and argued that the social product must be attributed entirely to acts of labour, capital being merely the embodied labour of the past. On this view, profit and interest are the result of pure exploitation in the sense that they consist of an income derived from the power position of the capitalist and not from the performance of any service. Non-Marxist economists have generally followed Senior in finding some function in society that corresponds to these forms of income.

The Marginalists generally held that profit and interest were related to the marginal productivity of the extension of the period of production. Böhm-Bawerk assumed that "roundabout" processes of production would generally be more productive than processes with shorter periods of production; he thought there was a productivity of "waiting" (to use the term of Alfred Marshall) and saw the rate of interest as an inducement to the capitalist to extend the period of production.

A low rate of interest leads to concentration on longer, more roundabout processes, and a high rate of interest on shorter, less roundabout processes. There is a limit, however, on the period of production imposed by the existing stock of accumulated capital. If one embarks on a long process with insufficient capital, he will find that he has exhausted his resources before the end of the process and before the fruits can be gathered. It is the business of the rate of interest to prevent this, and to adjust the roundaboutness of the processes used to the capital resources available. The Marginalists' theory of interest reached its clearest expression in the work of Irving Fisher. He saw an equilibrium rate of interest as determined by the interaction of two sets of forces: the impatience of consumers on the one hand, and the returns from extending the period of production on the other.

John Maynard Keynes brought a new approach. His liquidity preference theory of interest is a short-run theory of the price of contractual obligations ("bonds"), and it is essentially an application of the general theory of market price. If people as a whole decide that they want to hold a larger proportion of their assets in the form of money, and if new money is not created to satisfy this desire, there will be a net desire to sell securities and the price of securities will fall. This is the same thing as a rise in the rate of interest. Conversely, if people want to get rid of money the price of securities will rise and the rate of interest will fall. This, then, is the theory of the "market" rate of interest, by contrast with the Marginalists' theory, which concerns itself with whether or not there is a long-run equilibrium rate of interest. The controversy, therefore, between the liquidity preference theory—which regards interest as a "bribe" to prevent people holding money rather than bonds—and the time preference theory—which regards interest as a bribe to persuade people to postpone enjoyments to the future—can be resolved by placing the former in the short run and the latter in the long run.

Contemporary questions. The middle of the 20th century saw a considerable shift in the focus of concern relating to the theory of interest. Economists seemed to lose interest in the equilibrium theory, and their main concern was with the effect of rates of interest as a part of monetary policy in the control of inflation. It was recognized that the monetary authority could control the rate of interest in the short run. The controversy lay mainly between the advocates of "monetary policy" and the advocates of "fiscal policy." If inflation is regarded as a symptom of a desire on the part of a society to consume and invest more in total than its resources permit, it is clear that the problem can be attacked either by diminishing investment or by diminishing consumption. On the whole, the attack of the advocates of monetary policy is on the side of diminishing investment, through raising rates of interest and making it harder to obtain loans, though the possibility that high rates of interest may restrict consumption is not overlooked. The alternative would seem to restrict consumption by raising taxes. This has the disadvantage of being politically unpopular. The mounting concern with economic growth, however, has raised considerable doubts about the use of high rates of interest as an instrument to control inflation. There is some doubt whether high interest rates in fact restrict investment; if they do not, they are ineffective, and if they do, they may be harmful to economic growth. This is a serious dilemma for the advocates of monetary policy. On the other hand, it must be admitted that the type of fiscal policy that might be most desirable theoretically has achieved very limited public support.

Interest rates and inflation

The ethics of interest

The problem of the ethics of interest is still unresolved after many centuries of discussion; as long as the institution of private property is accepted, the usefulness of borrowing and lending can hardly be denied. In the long historic process of inheritance, widowhood, gain and loss, by which the distribution of the ownership of capital is determined, there is no reason to suppose that the actual ownership of capital falls into the hands of those best able to administer it. Much of the capital of an advanced society, in fact, tends to be owned by elderly widows, simply because of the greater longevity of the female. Society, therefore, needs some machinery for separating the control of capital from its ownership. Financial instruments and financial markets are the principal agency for

performing this function. If all securities took the form of stocks or equities, it might be argued that contractual obligations (bonds), and therefore interest as a form of income, would not be necessary. The case for bonds and interest, however, is the case for specialization. There is a demand for many different degrees of ownership and responsibility, and interest-bearing obligations tap a market that would be hard to reach with equity securities; they are also peculiarly well adapted to the obligations of governments. The principal justification for interest and interest-bearing securities is that they provide an easy and convenient way for skilled administrators to control capital that they do not own and for the owners of capital to relinquish its control. The price society pays for this arrangement is interest.

There remains the problem of the socially optimum rate of interest. It could be argued that there is no point in paying any higher price than one needs to and that the rate of interest should be as low as is consistent with the performance of the function of the financial markets. This position, of course, would place all the burden of control of economic fluctuations on the fiscal system, and it is questionable whether this would be acceptable politically.

The ancient problem of "usury," in the form of the exploitation of the ignorant poor by moneylenders, is still important in many parts of the world. The remedy is the development of adequate financial institutions for the needs of all classes of people rather than the attempt to prohibit or even to limit the taking of interest. The complex structure of lending institutions in a developed society—banks, building societies, land banks, cooperative banks, credit unions, and so on—testifies to the reality of the service that the lender provides and that interest pays for. The democratization of credit—that is, the extension of the power of borrowing to all classes in society—is one of the important social movements of the 20th century.

(K.E.Bo.)

LABOUR AND WAGES

Wages are income derived from human labour. Technically they cover all payments for the use of labour, mental or physical, but in ordinary usage the term excludes income of the self-employed and is restricted to compensation of employees. Occasionally fringe benefits are included, but generally they are not. The term is not fully synonymous with labour costs, which may include such items as cafeterias or meeting rooms maintained for the convenience of employees (such items are part of capital). Wages, in economic terms, however, do include remuneration in the form of extra benefits, such as paid vacations, holidays, and sick leave, as well as wage supplements in the form of pensions and health insurance paid for by the employer. A worker in covered industries also receives the protection of governmentally provided unemployment compensation, old-age pensions, and industrial accident compensation. Government services provided for workers are of even greater significance in European countries than in the United States and must be taken into account when comparisons of earnings are made.

Wages and labour costs

Classical theories. Theories of wage determination and the share of labour in the gross national product have varied from time to time and have changed as the economic environment has changed. The body of thought referred to today as wage theories could not have emerged until the old feudal system had disappeared and the modern economy with its modern institutions had come into existence. Adam Smith, in *The Wealth of Nations* (1776), failed to propose a definitive theory of wages, but he anticipated several theories that were developed by others later. Smith thought that wages were determined in the marketplace through the law of supply and demand. Workers and employers would naturally follow their own self-interest; labour would be attracted to the jobs where labour was needed most, and the result would be the greatest overall benefit to the workers and to society. But Smith gave no precise analysis of the supply of and demand for labour; he discussed many elements that were involved but did not weave them into a consistent theoretical pattern.

Subsistence theory. Subsistence theories emphasize the supply aspects and neglect the demand aspects of the labour market. They hold that change in the supply of workers is the basic force that drives real wages to the minimum required for subsistence. Elements of a subsistence theory appear in *The Wealth of Nations*, where Smith wrote that the wages paid to workers had to be enough to allow them to live and to reproduce themselves. Smith was more optimistic, however, than the British classical economists, such as David Ricardo and Thomas Malthus, who followed him, for he implied that—at least in an advancing nation—the wage level would have to be above subsistence to permit the population to grow enough to supply the additional workers needed. Ricardo maintained a more rigid view. He wrote that the "natural price" of labour was the price necessary to enable the labourers to subsist and to perpetuate the race without increase or diminution. Ricardo's statement was consistent with the Malthusian theory of population, which held that population adjusts to the means of supporting it. The market price of labour could not vary from the natural price for long: if wages rose above subsistence, the number of workers would increase and bring the wage rates down; if wages fell below subsistence, the number of workers would decrease and bring the wage rates up. At the time that these economists wrote, most workers were actually living near the subsistence level, and population appeared to be trying to outrun the means of subsistence. The subsistence theory seemed to fit the facts; and, although Ricardo said that the natural price of labour was not fixed and might be changed if custom and habit moderated population increases in relation to food supply and other items necessary to maintain labour, later writers tended to subscribe to the basic idea and not to admit exceptions. Their inflexible and inevitable conclusion earned the theory the name "iron law of wages."

The "natural price" of labour

Wages-fund theory. Smith said that the demand for labour could not increase except in proportion to the increase of the funds destined for the payment of wages. Ricardo maintained that an increase in capital would result in an increase in demand for labour. Statements such as these foreshadowed the wages-fund theory, which held that a predetermined fund of wealth existed for the payment of wages. The size of the fund could be changed over periods of time, but at any given moment the amount was fixed, and the average wage could be determined simply by dividing the fund by the number of workers. Smith thought of the fund as surplus income of wealthy men—beyond the needs of their families and trade—which they would use to employ others. Ricardo thought of it in terms of capital—food, clothing, tools, raw materials, machinery, etc., necessary to give effect to labour. Regardless of the makeup of the fund, the obvious conclusion was that when the fund was large in relation to the number of workers, wages would be high. When it was relatively small, wages would be low. If population increased too rapidly in relation to food and other necessities (as outlined by Malthus), wages would be driven to the subsistence level. Therefore, it would be to the advantage of labour to help promote the accumulation of capital to enlarge the fund rather than to discourage it by forming labour organizations and making exorbitant demands. Also, it followed that legislation designed to raise wages would not be successful, for, with only a fixed fund to draw upon, increases gained by some workers could be maintained only at the expense of others.

This theory was generally accepted for 50 years by economists, including such well-known figures as Nassau William Senior and John Stuart Mill. W.T. Thornton, F.D. Longe, and Francis A. Walker were largely responsible for discrediting the theory during the decade following 1865. They pointed out that the demand for labour was not determined by a fund but was derived from the consumer demand for products. The proponents of the wages-fund doctrine had been unable to prove that there was a determinate wage fund, or any fund maintaining a predetermined relationship with capital or with the portion of the proceeds of labour's product paid out in wages. Actually the amount paid out depended upon a number of factors, including the bargaining power of labour. Yet,

in spite of these telling criticisms, the wages-fund theory continued to exercise an important influence until the end of the 19th century.

Marxian surplus-value theory. Karl Marx accepted Ricardo's labour theory of value, but he subscribed to a subsistence theory of wages for a different reason than that given by the classical economists. In Marx's mind, it was not the pressure of population that drove wages to the subsistence level but rather the existence of a large army of unemployed, which he blamed on the capitalists. He stated that the exchange value of any product was determined by the amount of labour time socially necessary to create it. He held that under the capitalistic system, labour was merely a commodity and could get only its subsistence. The capitalist, however, could force the worker to spend more time on his job than was necessary to earn his subsistence, and the excess product, or surplus value, thus created, was taken by the capitalist.

Product in excess of subsistence

From the point of view of classical theory, Marx's argument appeared persuasive, although the term "labour time socially necessary" hid some serious objections. The fatal blow came when the labour theory of value and Marx's subsistence theory of wages were found to be invalid. Without them, the surplus-value theory collapsed.

Residual-claimant theory. The residual-claimant theory holds that, after all other factors of production have received their share of the product, the amount left goes to the remaining factor. Adam Smith implied such a theory for wages, since he said that rent would be deducted first and profits next. Francis A. Walker in 1875 worked out a residual theory of wages in which the shares of the landlord, capitalist, and entrepreneur were determined independently and subtracted, thus leaving the remainder for labour in the form of wages. It should be noted, however, that any of the factors of production may be selected as the residual claimant, assuming that independent determinations may be made for the shares of the other factors. It is doubtful, therefore, that such a theory has much value as an explanation of wage phenomena.

Bargaining theory. The bargaining theory of wages holds that wages, hours, and working conditions are determined by the relative bargaining strength of the parties to the agreement. Smith hinted at such a theory when he noted that employers had greater bargaining strength than employees, because it was easier for employers to combine in opposition to employees' demands and also because employers were financially able to withstand the loss of income for a longer period than the employees. This idea was developed to a considerable extent by John Davidson, who argued, in 1898, that the determination of wages is an extremely complicated process involving numerous influences that interact to establish the relative bargaining strength of the parties. There is no one factor or single combination of factors that determines wages, and there is no one rate that necessarily prevails. Because there are many possibilities, there is a range of rates within which any number of rates may exist simultaneously. The upper limit of the range is set by the rate beyond which the employer refuses to hire certain workers. This rate is influenced by such considerations as the productivity of the workers, the competitive situation, the size of the investment, and the employer's estimate of future business conditions. The lower limit of the range is set by the rate below which the workers will not offer their services to the employer. This rate is influenced by such considerations as minimum wage legislation, the workers' standard of living, their appraisal of the employment situation, and their knowledge of rates paid to others. Neither the upper nor the lower limit is fixed, and either may move upward or downward. The rate or rates within the range are determined by relative bargaining power.

The bargaining theory is very attractive to labour organizations, for, contrary to the subsistence and wages-fund theories, it provides a very cogent reason for the existence of unions. The bargaining strength of a union is much greater than that of the members acting as individuals. Also there are situations (bilateral monopoly, for instance) under which theoretical analysis arrives at a range of wage rates rather than a determinate rate. The actual rate must depend upon relative bargaining power. It should be observed, however, that historically labour was able to improve its situation before its bargaining power became more effective through organization. Factors other than the relative bargaining strength of the parties must have been at work. The bargaining theory often gives an excellent explanation of a short-run situation, such as the existence of certain wage differentials, but over the long run it fails to provide an adequate understanding of the changes that have taken place in the average level of wages.

Short-run and long-run applications

Marginal-productivity theory and its critics. Toward the end of the 19th century, marginal-productivity analysis was applied not only to labour but to other factors of production as well. It was not a new idea as an explanation of wage phenomena, for Smith had observed that a relationship existed between wage rates and the productivity of labour, and Johann Heinrich von Thünen, a German economist, had worked out a marginal-productivity type of analysis for wages in 1826. The Austrian economists made important contributions to the marginal idea after 1870; and, building on these grounds, a number of economists in the 1890s, including Philip Henry Wicksteed in England and John Bates Clark in the United States, elaborated the idea into the marginal-productivity theory of distribution. It is likely that the disturbing conclusions drawn by Marx from classical economic theory inspired this development. In the early 1930s refinements to the marginal-productivity analysis, particularly in the area of monopolistic competition, were made by Joan Robinson in England and Edward H. Chamberlin in the United States.

As applied to wages, the marginal-productivity theory holds that employers will tend to hire workers of a particular type until the addition made by the last (marginal) worker to the total value of the product is equal to the addition to total cost caused by the hiring of one more worker. The wage rate is established in the market through the demand for, and supply of, the type of labour, and the operation of competition assures the workers that they will receive a wage equal to the marginal product. Under the law of diminishing marginal productivity, the contribution of each additional worker is less than that of his predecessor, but workers of a particular type are assumed to be alike, making them interchangeable, and any one could be considered the marginal worker. All receive the same wage, and, therefore, by hiring to the margin, the employer maximizes his profits. As long as each additional worker contributes more to total value than he costs in wages, it pays the employer to continue hiring. Beyond the margin, additional workers would cost more than their contribution and would subtract from attainable profits.

The theory also provides an explanation of wage differentials. Wage differentials are caused by differences in marginal product. The wages of skilled workers are higher than those of unskilled workers because there are fewer skilled workers, and their marginal product, therefore, is higher.

The marginal-productivity theory of wages became the prevailing wage theory, and, although it has been attacked by many and discarded by some, no acceptable alternative has been devised. The chief basis for criticism of the theory is that it rests on unrealistic assumptions, such as the existence of homogenous groups of workers whose knowledge of the labour market is so complete that they will always move to the best job opportunities. Workers are, in fact, not homogenous; usually they have little knowledge of the labour market; and because of home ties, seniority, and other considerations, they do not often move quickly from one job to another. The assumption that employers are able to measure productivity accurately and compete freely in the labour market also is farfetched. Even the assumption that all employers attempt to maximize profits may be doubted. The profit motive does not affect charitable institutions or government agencies. For the theory to operate properly, labour and capital must be fully employed so that increased production can be secured only at increased cost; capital and labour must be easily substitutable for each other; and the situation must be completely competitive.

Theoretical weaknesses

Obviously these assumptions do not fit the real world, and some critics feel that the results of the theory are so misleading that the theory should be abandoned. The proponents argue, however, that productivity gives a rough approximation of wages, and that although productivity may not provide the immediate explanation in a particular case, it certainly indicates long-run trends. The theory, therefore, has important uses, and if the difficulties are kept in mind, it can be a valuable tool.

In a modern economy, monopolistic or near monopolistic conditions exist in some important areas, particularly where there are only a few large producers (such as in the automobile industry) on one side of the bargaining table and powerful labour organizations on the other. Under such circumstances, the marginal productivity analysis cannot determine wages precisely; it can show only the positions that the union (as a monopolist of labour supply) and the employer (as a monopsonist, or single purchaser of labour services) will strive to reach, depending upon their current policies.

Wages, consumption and employment

Purchasing-power theory. The purchasing-power theory of wages involves the relation between wages and employment and the business cycle and is not, therefore, a theory of wage determination. It stresses the importance of spending through consumption and investment as an influence upon the activity of the economy. The theory gained prominence during the Great Depression of the 1930s, when it became apparent that lowering wages might not increase employment as previously had been assumed. John Maynard Keynes, the British economist, maintained in his *General Theory of Employment, Interest and Money* (1936), that (1) depressional unemployment could not be explained merely by frictions in the labour market that interrupted the smooth movement of the economy toward full employment equilibrium and (2) the assumption that "all other things remained equal" presented a special case that had no real applicability to the existing situation. Keynes related changes in employment to changes in consumption and investment, and he pointed out that stable equilibrium could exist with less than full employment.

Because wages make up such a large percentage of the national income, changes in wages usually have an important effect upon consumption. It is possible that lowering wages will reduce consumption and that, with the decline in demand for goods and services, the demand for labour may also fall, thus decreasing employment rather than improving it. Whether this will be the result, however, depends upon several considerations, particularly the reaction upon prices. If wages fall more rapidly than prices, labour's real wages will be drastically reduced, and consumption will fall, accompanied by increased unemployment, unless total spending is maintained by increased investment. Entrepreneurs may look upon the lower wage costs in relation to prices as an encouraging sign toward greater profits, in which case they may increase their investments and employ more people at the lower rates, thus maintaining or even increasing total spending and employment. If employers look upon the falling wages and prices as an indication of further declines, however, they may contract their investments or do no more than maintain them. In this case, total spending and employment will decline.

If wages fall less rapidly than prices, labour's real wages will increase, and consumption may rise. If investment is at least maintained, total spending in terms of constant dollars will increase, thus improving employment. If entrepreneurs look upon the shrinking profit margin as a danger signal, however, they may reduce their investments; and, if the result is a reduction in total spending, employment will fall. If wages and prices fall the same amount, there should be no change in consumption and investment; and, in that case, employment will remain unchanged.

The purchasing-power theory involves psychological considerations as well as those that may be measured more objectively. Whether it can be used effectively to control the business cycle depends upon political as well as economic factors, because government expenditures are a part of total spending, taxes may affect private spending, etc. The applicability of the theory is to the whole economy rather than to the individual firm. (P.L.KI.)

LAND AND RENT

Rent in economics is specially defined. According to the neoclassical economist Alfred Marshall, rent is the income derived from the ownership of land and other free gifts of nature. He, and others after him, chose this definition for technical reasons, even though it is somewhat more restrictive than the meaning given the term in popular usage. Apart from renting land, it is of course possible to rent (in other words, to pay money for the temporary use of any property) houses, automobiles, television sets, and lawn mowers on the understanding that the rented item is to be returned to its owner in essentially the same physical condition.

The more restrictive use of the term became popular rather early among writers on economic matters. For the classical economists of the 18th and 19th centuries, society was divided into three groups: landlords, labourers, and businessmen (or the "moneyed classes"). This division reflected more or less the sociopolitical structure of Great Britain at the time. The concern of economic theorists was to explain what determined the share of each class in the national product. The income received by landlords as owners of land was called rent.

It was observed that the demand for the product of land would make it profitable to extend cultivation to soils of lesser and lesser fertility, as long as the addition to the value of output would cover the costs of cultivation on the least fertile acreage cultivated. On land of greater fertility—intramarginal land—the costs of cultivation per unit of output would be below that price. This difference between cost and price could be appropriated by the owners of land, who benefitted in this way from the fertility of the soil—a "free gift of nature."

Marginal land (the least fertile cultivated) earned no rent. Since, therefore, it was differences in fertility that brought about the surplus for landowners, the return to them was called differential rent. It was also observed, however, that rent emerged not only as cultivation was pushed to the "extensi ve margin" (to less fertile acreage) but also as it was pushed to the "intensive margin" through more intensive use of the more fertile land. As long as the additional cost of cultivation was less than the addition to the value of the product, it paid to apply more labour and capital to any given piece of land until the net value of the output of the last unit of labour and capital hired had fallen to the level of its incremental cost. The intensive margin would exist even if all land were of equal fertility, as long as land was in scarce supply. It can be called scarcity rent, therefore, to contrast it with differential rent.

Extensive and intensive margins of production

However, because the return to any factor of production, not only to land, can be determined in the same way as scarcity rent, it was often asked why the return to land should be given a special name and special treatment. A justification was found in the fact that land, unlike other factors of production, cannot be reproduced. Its supply is fixed no matter what its price. Its supply price is effectively zero. By contrast, the supply of labour or capital is responsive to the price that is offered for it. With this in mind, rent was redefined as the return to any factor of production over and above its supply price.

With the supply price of land zero, the whole of its return is rent, so defined. The return to any other factor may also contain elements of rent, as long as the return stands above the next-most-lucrative employment open to that factor. For example, a singer's employment outside the opera may bring a great deal less than the opera actually pays. A large part of what the opera pays must therefore be called rent.

The opera singer's specific talent may be nonreproducible; like land, it is a "free gift of nature." A particularly effective machine also, though its supply can be increased in time by productive effort, may for a period also earn a quasi-rent, until supply has caught up with demand. Where its supply is artificially restricted by a monopoly, the quasi-rent may in fact continue indefinitely. All monopoly profits, it has been argued, should therefore be classified as

quasi-rent. Once this point has been reached in the argument, there is perhaps no logical barrier to extending the meaning of rent to cover all property returns. After all, profits and interest can persist only as long as there is no glut of capital. The possibility of producing capital would presage such a glut, one that has been staved off only by new scarcities created by technical progress.

(H.O.Sc.)

GENERAL THEORIES OF DISTRIBUTION

The division of the social product

The theory of distribution deals with the way in which a society's product is distributed among the members of that society. It involves three distinguishable sets of questions. First, how is the national income distributed among persons? How many persons earn less than $10,000, how many between $10,000 and $20,000, how many between $20,000 and $30,000, and so on? Are there regularities in these statistics? Is it possible to generalize about them? This is the problem of personal distribution. Second, what determines the prices of the factors of production? What are the influences governing the wage rate for a specific kind of labour? Why is the general wage level of a country not lower or higher than it is? What determines the rate of interest? What determines profits and rents? These questions have to do with functional distribution. Third, how is the national income distributed proportionally among the factors of production? What determines the share of labour in the national income, the share of capital, the share of land? This is the problem of distributive shares. Although the three sets of problems are obviously interrelated, they should not be confused with one another. The theoretical approaches to each of them involve quite different considerations.

Aspects of distribution. *Personal distribution.* Personal distribution is primarily a matter of statistics and the conclusions that can be drawn from them. When incomes are charted according to the number of people in each size category, the resulting frequency distribution is rather startling. Generally the top 10 percent of income receivers get between 25 and 35 percent of the national income, while the lowest 20 percent of the income receivers get about 5 percent of the national income. The inequality seems to be greatest in poor countries and diminishes somewhat in the course of economic development.

There are various explanations of the inequality. Some authorities point to the natural inequality of human beings (differences in intelligence and ability), others to the effects of social institutions (including education); some emphasize economic factors such as scarcity; others invoke political concepts such as power, exploitation, or the structure of society.

Distributive shares

For a long time economists were pessimistic as to the possibilities of any substantial improvement in the lot of those at the bottom of the income distribution. They generally held that the scarcity of productive land and the tendency of population to increase faster than the means of subsistence imposed limits on distributive justice. David Ricardo, in *On the Principles of Political Economy and Taxation* (1817), held that the landlords would receive an increasing part of the national income while capitalists would get less and less and that this shift in distribution would lead to economic stagnation. Karl Marx prophesied that the workers would be increasingly exploited and made miserable and that these conditions would lead to the downfall of capitalism. Neither prediction materialized. Thus in the Western world the share of rents has dwindled to a few percent of the national income, while the share of labour has gradually increased. For some time, economists believed that the share of labour was more or less constant, but investigations show that economic development is accompanied by an increasing share of labour. Though the statistics are complicated by technical problems, it is safe to say that in the United States, the share of wages rose from more than half the national income at the beginning of the century to more than 70 percent in the 1980s.

Contemporary approaches to this aspect of income distribution vary. Some are highly abstract and are closely related to the study of the whole, the modern macroeconomics of saving and investment. These will not be dealt with here. A simple common-sense approach employs an equation that starts by writing labour's share as the quotient of the total wage bill and the national income, $\text{labour's share} = \frac{\text{total wage bill}}{\text{national income}}$, and then writes the wage bill as the product of the wage level and the amount of labour (wage bill = wage level × amount of labour); next the national income is written as the product of the national output and the price level (national income = national output × price level); the result is that the share of labour equals the quotient of the average real wage rate and labour productivity, $\text{share of labour} = \frac{\text{average real wage rate}}{\text{labour productivity}}$, the latter being the quotient of the national output and the amount of labour: $\text{labour productivity} = \frac{\text{national output}}{\text{amount of labour}}$. If these two variables move in a parallel fashion, the share of labour is constant. If the real wage rate increases faster than the amount of labour productivity, the share of labour goes up. Similar reasoning applies to the shares of capital and land. This simple arithmetic is useful for an understanding of what happens in the real world, but for a profounder analysis one must turn to the theory of functional distribution.

Functional distribution. The theory of functional distribution, which attempts to explain the prices of land, labour, and capital, is a standard subject in economics. It sees the demand for land, labour, and capital as derived demand, stemming from the demand for final goods. Behind this lies the idea that a businessman demands inputs of land, labour, and capital because he needs them in the production of goods that he sells. The theory of distribution is thus related to the theory of production, one of the well-developed subjects of economics. The reasoning that synthesizes production and distribution theory is called neoclassical theory.

The theory of marginal productivity

Components of the neoclassical, or marginalist, theory. The basic idea in neoclassical distribution theory is that incomes are earned in the production of goods and services and that the value of the productive factor reflects its contribution to the total product. Though this fundamental truth was already recognized at the beginning of the 19th century (by the French economist J.B. Say, for instance), its development was impeded by the difficulty of separating the contributions of the various inputs. To a degree they are all necessary for the final result: without labour there will be no product at all, and without capital total output will be minimal. This difficulty was solved by J.B. Clark (*c.* 1900) with his theory of marginal products. The marginal product of an input, say labour, is defined as the extra output that results from adding one unit of the input to the existing combination of productive factors. Clark pointed out that in an optimum situation the wage rate would equal the marginal product of labour, while the rate of interest would equal the marginal product of capital. The mechanism tending to produce this optimum begins with the profit-maximizing businessman, who will hire more labour when the wage rate is less than the marginal product of additional workers and who will employ more capital when the rate of interest is lower than the marginal product of capital. In this view, the value of the final output is separated (imputed) by the marginal products, which can also be interpreted as the productive contributions of the various inputs. The prices of the factors of production are determined by supply and demand, while the demand for a factor is derived from the demand of the final good it helps to produce. The word derived has a special significance since in mathematics the term refers to the curvature of a function, and indeed the marginal product is the (partial) derivative of the production function.

One of the great advantages of the neoclassical, or marginalist, theory of distribution is that it treats wages, interest, and land rents in the same way, unlike the older theories that gave diverging explanations. (Profits, however, do not fit so smoothly into the neoclassical system.) A second advantage of the neoclassical theory

is its integration with the theory of production. A third advantage lies in its elegance: the neoclassical theory of distributive shares lends itself to a relatively simple mathematical statement.

An illustration of the mathematics is as follows. Suppose that the production function (the relation between all hypothetical combinations of land, labour, and capital on the one hand and total output on the other) is given as $Q = f(L,K)$ in which Q stands for total output, L for the amount of labour employed, and K for the stock of capital goods. Land is subsumed under capital, to keep things as simple as possible. According to the marginal productivity theory, the wage rate is equal to the partial derivative of the production function, or $\partial Q/\partial L$. The total wage bill is $(\partial Q/\partial L) \cdot L$. The distributive share of wages equals $(L/Q) \cdot (\partial Q/\partial L)$. In the same way the share of capital equals $(K/Q) \cdot (\partial Q/\partial K)$. Thus the distribution of the national income among labour and capital is fully determined by three sets of data: the amount of capital, the amount of labour, and the production function. On closer inspection the magnitude $(L/Q) \cdot (\partial Q/\partial L)$, which can also be written $(\partial Q/Q)/(\partial L/L)$, reflects the percentage increase in production resulting from the addition of 1 percent to the amount of labour employed. This magnitude is called the elasticity of production with respect to labour. In the same way the share of capital equals the elasticity of production with respect to capital. Distributive shares are, in this view, uniquely determined by technical data. If an additional 1 percent of labour adds 0.75 percent to total output, labour's share will be 75 percent of the national income. This proposition is very challenging, if only because it looks upon income distribution as independent of trade union action, labour legislation, collective bargaining, and the social system in general. Obviously such a theory cannot explain all of the real economic world. Yet its logical structure is admirable. What remains to be seen is the degree to which it can be used as an instrument for understanding the real economic world.

Application of the theory to reality

Criticisms of the neoclassical theory. *Returns to scale.* Neoclassical theory assumes that the total product Q is exactly exhausted when the factors of production have received their marginal products; this is written symbolically as $Q = (\partial Q/\partial L) \cdot L + (\partial Q/\partial K) \cdot K$. This relationship is only true if the production function satisfies the condition that when L and K are multiplied by a given constant then Q will increase correspondingly. In economics this is known as constant returns to scale. If an increase in the scale of production were to increase overall productivity, there would be too little product to remunerate all factors according to their marginal productivities; likewise, under diminishing returns to scale, the product would be more than enough to remunerate all factors according to their marginal productivities.

Research has indicated that for countries as a whole the assumption of constant returns to scale is not unrealistic. For particular industries, however, it does not hold; in some cases increasing returns can be expected, and in others decreasing returns. This situation means that the neoclassical theory furnishes at best only a rough explanation of reality.

One difficulty in assessing the realism of the neoclassical theory lies in the definition and measurement of labour, capital, and land, more specifically in the problem of assessing differences in quality. In macroeconomic reasoning one usually deals with the labour force as a whole, irrespective of the skills of the workers, and to do so leaves enormous statistical discrepancies. The ideal solution is to take every kind and quality of labour as a separate productive factor, and likewise with capital. When the historical development of production is analyzed it must be concluded that by far the greater part of the growth in output is attributable not to the growth of labour and capital as such but to improvements in their quality. The stock of capital goods is now often seen as consisting, like wine, of vintages, each with its own productivity. The fact that a good deal of production growth stems from improvements in the quality of the productive inputs leads to considerable flexibility in the distribution of the national income. It also helps to explain the existence of profits.

Substitution problems. Another difficulty arises from the fact that marginal productivity assumes that the factors of production can be added to each other in small quantities. If one must choose between adding one big machine or none at all to production, the concept of the marginal product becomes unworkable. This "lumpiness" creates an indeterminacy in the distribution of income. From the viewpoint of the individual firm, this objection to neoclassical theory is more serious than from the macroeconomic viewpoint since in terms of the national economy almost all additions to labour and capital are very small. A related problem is that of substitution among factors. The production function implies that land, labour, and capital can be combined in varying proportions, that every conceivable input mix is possible. But in some cases the input mix is fixed (*e.g.*, one operator at one machine), and in that situation the neoclassical theory breaks down completely because the marginal product for every factor is zero. These cases of fixed proportions are scarce, however, and from a macroeconomic viewpoint it is safe to say that a flexible input mix is the rule.

This is not to say that substitution between labour and capital is so flexible in the national economy that it can be assumed that a 1 percent increase in the wage rate will reduce employment by a corresponding 1 percent. That would follow from the neoclassical theory described above. It is not impossible, but it requires a very special form of the production function known as the Cobb–Douglas function. The pioneering research of Paul H. Douglas and Charles W. Cobb in the 1930s seemed to confirm the rough equality between production elasticities and distributive shares, but that conclusion was later questioned; in particular the assumption of easy substitution of labour and capital seems unrealistic in the light of research by Robert M. Solow and others. These investigators employ a production function in which labour and capital can replace each other but not as readily as in the Cobb–Douglas function, a change that has two very important consequences. First, the effect of a wage increase on the share of labour is not completely offset by changes in the input mix, so that an increase in wage rates does not lead to a proportionate reduction in total employment; and second, the factor of production that grows fastest will see its share in the national income diminished. The latter discovery, made by J.R. Hicks (1932), is extremely significant. It explains why the remuneration of capital (interest, not profits) has shrunk from 20 percent or more a century ago to less than 10 percent of the national income in modern times. In a society where more and more capital is employed in production, a continually smaller proportion of the income goes to the owners of capital. The share of labour has gone up; the share of land has gone down dramatically; the share of capital has gradually declined; and the share of profits has remained about the same. This picture of the historical development of income distribution fits roughly into the frame of neoclassical theory, although one must also make allowance for the short-run effects of inflation and the long-run effects of technological progress.

The analysis of earnings

Returns to the factors of production. The demand side of the markets for productive factors is explained in large degree by the theory of marginal productivity, but the supply side requires a separate explanation, which differs for land, labour, and capital.

Rent. The supply of land is unique in being rather inelastic; that is, an increase in rent does not necessarily increase the amount of available land. Landowners as a group receive what is left over after the other factors of production are paid. In this sense, rent is a residual, and a good deal of the history of the theory of distribution is concerned with the issue whether rent should be regarded as part of the cost of production or not (as in Ricardo's famous dictum that the price of corn is not high because of the rent of land but that land has a rent because the price of corn is high). But inelasticity of supply is not characteristic only of land; special kinds of labour and the size of the total labour force also tend to be unresponsive to variations in wages. The Ricardian issue, moreover, was important in the context of an agrarian society; it lacks

significance now, when land has so many different uses.

Wages. In analyzing the earnings of labour, it is necessary to take account of the imperfections of the labour market and the actions of trade unions. Imperfections in the market make for a certain amount of indeterminacy in which considerations of fairness, equity, and tradition play a part. These affect the structure of wages—*i.e.*, the relationships between wages for various kinds of labour and various skills. Therefore one cannot say that the income difference between a carpenter and a physician, or between a bank clerk and a truck driver, is completely determined by marginal productivity, although it is true that in the long run the wage structure is influenced by supply and demand.

Effects of labour unions

The role of the trade unions has been a subject of much debate. The naive view that unions can raise wages by their efforts irrespective of market forces is, of course, incorrect. In any particular industry, exaggerated wage claims may lead to a loss of employment; this is generally recognized by union leaders. The opposite view, that trade unions cannot influence wages at all (unless they alter the basic relationship between supply and demand for labour), is held by a number of economists with respect to the real wage level of the economy as a whole. They agree that unions may push up the money wage level, especially in a tight labour market, but argue that this will lead to higher prices and so the real wage rate for the economy as a whole will not be increased accordingly. These economists also point out that high wages tend to encourage substitution of capital for labour (the cornerstone of neoclassical theory). These factors do indeed operate to check the power of trade unions, although the extreme position that the unions have no power at all against the iron laws of the market system is untenable. It is safe to say that basic economic forces do far more to determine labour's share than do the policies of the unions. The main function of the unions lies rather in modifying the wage structure; they are able to raise the bargaining power of weak groups of workers and prevent them from lagging behind the others.

Interest and profit. The earnings of capital are determined by various factors. Capital stems from two sources: from saving (by households, financial institutions, and businesses) and from the creation of money by the banks. The creation of money depresses the rate of interest below what may be called its natural rate. At this lower rate, businessmen will invest more, the capital stock will increase, and the marginal productivity of capital will decline. Although this chain of reactions has drawn the attention of monetary theorists, its impact on income distribution is probably not very important, at least not in the long run. There are also other factors, such as government borrowing, that may affect the distribution of income; it is difficult to say in what direction. The basic and predominant determinant is marginal productivity: the continuous accumulation of capital depresses the rate of interest.

One type of earning that is not explained by the neoclassical theory of distribution is profit, a circumstance that is especially awkward because profits form a substantial part of national income (20–25 percent); they are an important incentive to production and risk taking as well as being an important source of funds for investment. The reason for the failure to explain profit lies in the essentially static character of the neoclassical theory and in its preoccupation with perfect competition. Under such assumptions, profit tends to disappear. In the real world, which is not static and where competition does not conform to the theoretical assumptions, profit may be explained by five causes. One is uncertainty. An essential characteristic of business enterprise is that not all future developments can be foreseen or insured against. Frank H. Knight (1921) introduced the distinction between risk, which can be insured for and thus treated as a regular cost of production, and uncertainty, which cannot. In a free enterprise economy, the willingness to cope with the uninsurable has to be remunerated, and thus it is a factor of production. A second way of accounting for profits is to explain them as a premium for introducing new technology or for producing more efficiently than one's competitors. This dynamic element in profits was stressed by Joseph Schumpeter (1911). In this view, prices are determined by the level of costs in the least progressive firms; the firm that introduces a new product or a new method will benefit from lower costs than its competitors. A third source of profits is monopoly and related forms of market power, whether deliberate as with cartels and other restrictive practices or arising from the industrial structure itself. Some economists have developed theories in which the main influence determining distributive shares is the relative "degree of monopoly" exerted by various factors of production, but this seems a bit one-sided. A fourth source of profits is sudden shifts in demand for a given product—so-called windfall profits, which may be accompanied by losses elsewhere. Finally, there are profits arising from general increases in total demand caused by a certain kind of inflationary process when costs, especially wages, lag behind rising prices. Such is not always the case in modern inflations.

Dynamic influences on distribution. *Prices.* Neoclassical theory throws light upon the long-run changes in distribution of income. It fails to take account of the short-run impact of business fluctuations, of inflation and deflation, of rapidly rising prices. This failure is an omission, though it is true that distributive shares do not fluctuate as much as employment, prices, and the state of business generally. This lagging in the behaviour of shares can be understood by remembering that they are determined by the quotient of the real remuneration of the factor and its productivity; both variables move, according to marginal productivity theory, in the same direction. Yet inflation and deflation do have a certain impact upon distribution: if purchasing power shrinks, profits are the first income category to suffer; next come wages, particularly through the effects of unemployment. In a depression, the recipients of fixed money incomes (such as interest and pensions) gain from lower prices. In an inflation the opposite happens.

How incomes are affected by changing conditions

The traditional inflationary sequence was that as prices rose, profits would increase, with wages lagging behind; this would tend to diminish the share of labour in the national income. Experience since World War II, however, has been different: in many countries wage levels tended to run ahead in the inflationary spiral and profits lagged behind, although most entrepreneurs eventually succeeded in shifting the burden of wage inflation onto the consumers. The result of the postwar inflation was a slight acceleration of the increase in the share of labour, while the shares of capital and land decreased faster than they would have in the absence of inflation. Profits as a whole held their own. The struggle among the various participants in the economic process no doubt added fuel to the inflationary fires.

Technology. Another dynamic influence is technological progress. The concept of the production function assumes a constant technology. But in reality the growth of production is much less the consequence of increased quantities of labour and capital than of improvements in their quality. This element in increased production is distributed in a way not fully explained by neoclassical theory. Part of the change in distribution that is caused by technological progress can be analyzed as resulting from changes in the elasticities of production. If $\frac{K}{Q} \cdot \frac{\partial Q}{\partial K}$ goes up, technological change is said to be "capital-using." and the share of capital will increase. This is what, in fact, may have happened; the change in technology has offset, though it has not neutralized, the decline in the share of capital caused by the employment of a higher amount of capital per worker. But another part of the fruits of technological progress is garnered by profit receivers, probably quite a substantial part. Businessmen who are quick innovators make high profits; in a rapidly changing society, profits tend to be high, a circumstance that is fortunate because profits are the mainspring of economic change. The high rate of growth experienced by the post-World War I Western world stemmed from this profit-innovation-profit nexus.

Personal income and neoclassical theory. The neoclassical theory endeavours to explain the prices of productive factors and the distributive shares received by them. It does not come to grips with a third category of distribution, that of personal income, which is much more affected by institutional arrangements and by characteristics of the social

structure. Profits in particular may be shared in various ways: they may accrue to stockholders, to workers, to management, or to the government; or they may be retained in the corporation. What happens depends on dividend policy, tax policy, and the existence of profit-sharing arrangements with workers. Neoclassical theory has little to say on these matters or on the fact that in present-day capitalist society the managers of big business are virtually in a position to fix their own personal incomes. Managers have so much power vis-à-vis the stockholders and their total share of profits is so relatively little that their ability to pay themselves high salaries is limited only by the conventions of the business world. These high incomes cannot be explained by the categories of the neoclassical theory, and they do not constitute an argument against the theory. They may well argue for changes in society's institutions, but that is a matter on which the neoclassical theory of distribution does not pontificate. A great deal of change could occur in the legal and social order without any disturbance to the theory. (J.P.)

Consumption

In economics the word consumption means the using up of goods and services. In modern economic terms it means, specifically, "final" consumption as distinguished from the using up of goods to produce other goods in a manufacturing industry. Final consumption must also be distinguished from the purchase by industry of fixed assets such as buildings and machinery, which is known as capital formation or investment. On the other hand, consumption expenditure by private persons is understood to include the purchase of durable goods, such as furniture or vehicles, as well as works of art that may increase in value over a period of time. The acquisition of such goods should actually be considered asset formation rather than consumption and should be classified with the acquisition of other assets such as houses, schools, roads, and hospitals.

Final consumption versus capital formation

In modern industrial economies, consumption as previously defined accounts for 70 or 80 percent of total national expenditure. Even in the Western capitalist countries a significant part of total consumption is determined directly by the expenditure of public authorities. Some of the benefits of this part of consumption, such as expenditure on defense or on public health, are widely diffused; others are directed by common consent to the benefit of particular sections of the community. These consist in part of specialized services such as education or medical care; but other services—such as unemployment compensation, state pensions for the elderly, and assistance to families deprived of the support of a wage earner—are designed to create greater equality in levels of consumption than would otherwise be obtained.

PATTERNS OF NATIONAL CONSUMPTION

The ways in which people spend their incomes show much uniformity among countries at the same economic level. Expenditure patterns in the United Kingdom, for example, are typical of western Europe. In 1949 the pattern was still affected by postwar shortages and rationing, but the level of total consumption was not very different from what it had been before the war. In the decades that followed, private consumption expenditure per person (measured at constant prices) doubled. In addition there was a great increase in public services such as health and education. Yet the broad distribution of expenditures remained strikingly constant in spite of the introduction of many new commodities and considerable changes in their relative prices. The percentage of total expenditure devoted to food fell, a phenomenon that usually accompanies a rising standard of living, and the largest proportionate increases were in the purchase and maintenance of private motor vehicles, of furniture and household goods, and of radio, television, and electrical goods. These three categories represent in part net additions to private wealth in the form of durable goods and also reflect the effect of technical progress. As in other industrial countries, much of the improvement in living standards has taken the form of more travel, better communication services, and the acquisition of labour-saving equipment.

Differences among countries

In most of the industrialized countries there has been a compound rate of increase in the total volume of consumption expenditure per person of 10 to 12 percent per year, the main exceptions to this being the United Kingdom and Japan, where consumption has grown at double this rate. But the pattern of change is similar in almost all countries. Food consumption has grown less rapidly than total consumption, particularly in the Scandinavian countries, Germany, and the countries of North America, where the rate of increase has been about 7 percent per year; expenditure on clothing has been growing at about the same rate as total consumption. Increases in rent outlays reflect higher energy costs in all of the industrialized countries. The acquisition of durable goods continues at a very high rate in all countries.

Comparable data on consumption in the poorer countries of the world are much harder to obtain and are usually less reliable, but it is probable that, expressed as a proportion of total consumption, food expenditure is about twice as important in much of Asia, Africa, and Latin America as it is in western Europe and North America. In the most economically advanced countries, food expenditure represents only one-quarter to one-third of the total, whereas in countries where the total expenditure per household is less than the equivalent of $1,500, the proportion rises to one-half or even greater. It should be noted that in the rural regions of poor countries the housing expenditure is minimal; in these areas shelter is rudimentary and largely self-provided.

Food consumption varies in character from country to country. This variation is due in part to climatic factors, and it also reflects differences in national food habits. The diet that is normally eaten in northern Europe and in Scandinavia is relatively low in fruit and vegetables but it contains a high proportion of milk, fats, and sugar. In France the consumption of vegetables and meat is relatively high. Fruit and vegetable consumption is generally high in southern Europe, while milk consumption in this area is low. In the Mediterranean countries food grains are generally preferred to potatoes and sugar as sources of carbohydrates.

But aside from these regional variations, the influence of general living standards is evident. The North American diet, for example, with its low grain and potato consumption and high consumption of sugar, meat, eggs, and fats is attributable more to a high standard of living than to any regional peculiarities of taste. These characteristics can be observed in the diets of the wealthier classes of most countries.

The influence of the general standard of living is also shown in the relative priorities that are accorded to the increased consumption of particular foods as incomes increase. These priorities are measured by economic statisticians in the form of income elasticities of expenditure, defined as the percentage increase in the consumption of an item divided by the percentage increase in income that makes the increased consumption possible. These elasticities are usually calculated for a given country by comparing the budgets of wealthy households with those of poor families. In countries such as the United States and Great Britain, the consumption of cereal foods actually decreases as incomes increase. In the less developed countries the elasticities are usually considerably higher, particularly for fruit and for products of animal origin. In these countries the consumption of carbohydrate foods is also increasing fairly rapidly as incomes rise.

THEORIES OF CONSUMER BEHAVIOUR

Factors influencing consumers. *Model of consumer behavior.* The theory and measurement of consumer behaviour forms an important part of modern economic theory. It was first developed during the 19th century on the basis of the following conceptions: that the purchase of any commodity gives the consumer a positive satisfaction or utility; the additional satisfaction derived from additional purchases of the same commodity declines as the consumer's supply of that commodity increases; and

with a given amount of money to spend, the consumer distributes the expenditure among commodities to maximize the total satisfaction or utility attainable from all those purchases. This rather crude model of consumer behaviour has undergone considerable refinement by modern mathematical economists. The advantage of this approach, which has had a strong and enduring effect on the theoretical and empirical work of economists, is that it separates the main economic variables influencing consumer behaviour—that is, income and prices—from all the remaining influences, such as individual preferences, social pressures, customs, and habits, but at the same time it unites them in a single analytical apparatus. Critics have often objected that the model assumes a rational person bent on scrupulously maximizing his satisfaction and that the model is thus part of a mechanistic stream of thought that has been substantially undermined by 20th-century advances in psychology. Still, the only useful criterion of any hypothesis is the range of situations in which the derivative model is shown validly to predict events. For example, it is useful to assume that the leaves of a tree attempt to maximize the amount of sunlight they receive, since the assumption implies that leaves are denser on the sunny side of trees than on the shady side, which can be checked from experience, or that billiard players make their shots as if they knew the mathematical formulas of mechanics. Similarly, to assume that consumers behave as if they were rational utility maximizers helps to provide accurate predictions of a broad range of market phenomena: *e.g.*, a fall in the price of a commodity will generally lead to increased consumption of that commodity, and an increase in consumer income will lead to increased consumption of most commodities. Only persistent discrepancies between predictions and events require a modification of the model's assumptions; some examples of such cases are discussed below.

Spending in relation to incomes

Income as determinant. The theory points to the income of consumers as the most important single determinant of their consumption patterns. It follows that in any community both the average income level and the distribution of incomes are important influences on total consumption. A community in which incomes are equally distributed consumes fewer luxury goods and fewer low-quality goods than one containing a few wealthy individuals and many poor people. Among wealthy people in early 19th-century England, a dinner with five main protein dishes—fish, meat, game, poultry, and ragout with truffles—was described as the minimum, while in poor years the families of agricultural labourers ate mainly oatmeal and potatoes; today the standardized produce of modern agriculture is part of most diets.

The classic model of consumers' behaviour implicitly assumes that the individual enjoys a constant income. In practice it may fluctuate according to the season, from year to year, or more generally over a lifetime. In the short run the consumption of some commodities is much affected by these income fluctuations, while the consumption of others is affected very little. Wage-earning households commonly have a weekly housekeeping allowance, out of which the necessities of food and clothing are bought, while the variable excess of earnings is spent on tobacco, alcoholic drinks, and entertainment. The expected average level of future income therefore influences consumption habits as much as actual present income, and commodities may be divided into two classes. The first consists of goods people buy when temporarily affluent but give up when temporarily poor, and the second consists of goods for which the pleasure of a temporarily higher level of consumption would not be worth the financial or psychological cost of giving them up in the future.

Consumers can also be influenced by their previous incomes. A person who owns an expensive car may continue to use it after his income falls, though at the lower level of income the individual would not choose to replace it with a similarly expensive vehicle in the long run. This may be a rational decision, in the sense that the value of the car in use may be greater than what it is worth in the second-hand automobile market; or it may be irrational, in the sense that an expensive habit that should have been abandoned is continued beyond the point where it can rationally be supported. The distinction is largely subjective and cannot be clearly made by an outside observer.

Over the life cycle as a whole, consumption patterns are markedly different in various occupations. In most of the unskilled or semiskilled occupations, the course of earnings is fairly stable: a young worker of 21 may earn as much as an older person. But in many of the professions an individual of 50 or 60 may earn many times the income of a person of 21; this gives the young wage earner a strong incentive to incur considerable debt with the expectation of amortizing it steadily throughout life, so that the typical consumption pattern of the occupation can be achieved earlier than otherwise. This applies particularly to such major purchases as houses, household furniture and equipment, and vehicles. Manual workers, on the other hand, whose expectations are little greater than their present consumption, generally prefer to rent living space rather than to own a house and are unwilling to raise their current consumption standards by incurring commitments of a longer term than that of ordinary installment credit.

Nonrational influences. To be fully rational and consistent, consumers need to have access to sufficient information on goods and their prices so that they can choose those with the lowest unit price for a given quality. But consumers do not always behave this way. Natural pearls are sold at a much higher price than cultured pearls, though the difference between them is demonstrable only by dissection or with X-rays, and their quality in use is identical. Brand-name drugs sell better and at higher prices than unbranded drugs that are manufactured from the same standard formula. To some extent this is due to what an American economist, Thorstein Veblen, called the desire for conspicuous consumption: part of the attraction of the good is simply its high price. It is also the result of consumers' ignorance, made more acute by the increasing sophistication of commodities whose qualities must be measured in many dimensions. If it is costly in time for the individual to become fully informed about the comparative qualities of competing products, it is not wholly irrational for the consumer to take the market price as an indicator of quality. The lack of information has given rise to consumers' organizations in most industrialized countries; these organizations test and report on a wide range of products for their subscribers.

Conspicuous consumption

The influence of modern advertising techniques must also be considered. Insofar as advertising informs the consumer of the range of alternatives, it can be argued that advertising merely increases the consumer's information; and insofar as advertising consciously or subconsciously changes consumer preferences, it remains one of the many factors determining consumer preferences that the economist takes as given. Advertising, however, cannot persuade the public to buy whatever the producer offers. Advertising is likely to be most effective in influencing consumers to choose one of several almost identical products being offered, such as toothpaste, cigarettes, or gasoline. But it may also raise the demand for the group of competing products as a whole. In addition, it can be argued that the total effect of modern advertising is to shift the preferences of consumers in favour of luxury goods rather than necessities, in favour of consumption rather than saving, and in favour of employment rather than leisure.

Attitudes toward necessities and luxuries. The distinction between necessities and luxuries is imprecise. The dividing line varies with the income and social class of the classifier and shifts as technology develops and as social values change. Only in the most undeveloped communities can necessities be defined purely in terms of physiological needs. Adam Smith wrote in 1776:

> By necessaries I understand not only the kind of commodities which are indispensably necessary for the support of life, but whatever the custom of the country renders it indecent for creditable people, even of the lowest order, to be without. . . . Under necessaries, therefore, I comprehend not only those things which nature, but those things which the established rules of decency have rendered necessary to the lowest rank

of people. All other things I call luxuries; without meaning by this appellation to throw the smallest degree of reproach upon the temperate use of them. Beer and ale, for example, in Great Britain, and wine, even in the wine countries, I call luxuries. A man of any rank may, without any reproach, abstain totally from tasting such liquors. Nature does not render them necessary for the support of life, and custom nowhere renders it indecent to live without them.

In the 19th century, with the development of more mathematical methods of reasoning based on a utilitarian calculus, the distinction came to be phrased differently. Necessities were defined as those commodities whose demand has an income elasticity less than unity, and luxuries as those with an income elasticity greater than unity. These definitions imply that as a worker's income increases the expenditure on necessities increases less than, and the expenditure on luxuries more than, proportionately. But even with the elasticity approach the distinction must vary over time. In 1950 the demand for television sets had a high income elasticity, whereas now, in some countries, television often is regarded as a necessity.

Changing attitudes toward luxury

Economists of the early 19th century all believed that the living standards of the working classes in capitalist societies would remain close to a subsistence level, meaning that luxuries would be more or less permanently denied them. But in modern industrialized economies even the poor consume goods that the early economists would not have considered necessary.

Role of luxuries. The historical and social role of luxury consumption is a subject of much interest. In the Mediterranean city-states during the Renaissance, the demand for luxuries provided a mainspring for the specialization of skilled labour and for the development of foreign travel and long-distance trade. The duke of Milan, Filippo Maria Visconti, possessed valuable English dogs, leopards from all parts of the East, and hunting birds from northern Europe. Some writers have argued that the luxurious consumption of the rich benefits the poor through the provision of employment opportunities that would not otherwise exist. A subtler version of this idea was proposed by Adam Smith, who contrasted the uselessness of menial labour employed by the rich for personal services with the benefits flowing from the employment of craftsmen who created luxurious products of enduring merit that eventually became available to society as a whole:

> The houses, the furniture, the clothing of the rich, in a little time, become useful to the inferior and middling ranks of people. . . . What was formerly a seat of the family of Seymour is now an inn upon the Bath road. The marriage-bed of James the First . . . was, a few years ago, the ornament of an ale-house at Dunfermline.

But Smith and most of the economists who succeeded him believed that, if the money spent on luxurious consumption by the rich was invested in useful production, society would benefit as a whole. The Industrial Revolution brought an increasing demand for funds for productive investment and made possible a more rapid rise in general standards of living than the world had known before. The classical economists thus argued that all luxury consumption involved a selfish diversion of labour and capital and acted as a brake on human progress.

This view was not seriously challenged until the English economist John Maynard Keynes published his *General Theory of Employment, Interest and Money* in 1935–36. Writing at a time when millions of workers were unemployed, Keynes argued that the consumption of luxuries was socially desirable if it provided jobs that would otherwise not exist. He also suggested that capitalism might be outrunning its investment opportunities, so that in the long run the problem of finding employment for capital itself would arise—a difficulty that might be postponed if the wealthy spent more on themselves:

> In so far as millionaires find their satisfaction in building mighty mansions to contain their bodies when alive and pyramids to shelter them after death, . . . the day when abundance of capital will interfere with abundance of output may be postponed.

In industrial countries since World War II, this pessimistic view has been overborne by a seemingly endless expansion in consumer industries. As fast as consumers accumulate durable goods, they become technologically or conventionally obsolete and are replaced by new goods. Instead of seeking more leisure, previously thought to be a main benefit of technical progress, the populations of the industrialized countries seem to prefer to work in order to buy more luxuries. To this extent the desire for leisure and the demand for luxuries are in direct competition.

Luxuries and leisure

Standards of consumption. In communist countries, public consumption has long been treated as more important than private luxury. In the last part of the 20th century this emphasis seemed to be giving way to the aim of catching up with the standards of consumption that prevail in capitalist countries. In the less developed countries of the Third World, the tension between the demand for luxuries and the low standard of living gives rise to acute economic and social problems. The rapid growth of international travel and communications since World War II has led the literate and skilled classes of every nation to seek similar standards of private consumption regardless of their national environment. This, in less developed countries, leads either to a highly unbalanced distribution of the national income or to the emigration of the skilled population. Thus the increasing awareness of the consumption habits of the most fortunate sections of the world's population is both a spur and a hindrance to general progress. (J.A.C.B.)

Economic fluctuations: stability and instability

BUSINESS CYCLES

Business cycles are best defined as fluctuations in the general level of economic activity, or more specifically, in the levels of employment, production, and prices. Figure 11 shows fluctuations in wholesale prices in four Western

Reprinted from A. Burns and W. Mitchell, *Measuring Business Cycles*; by permission of National Bureau of Economic Research

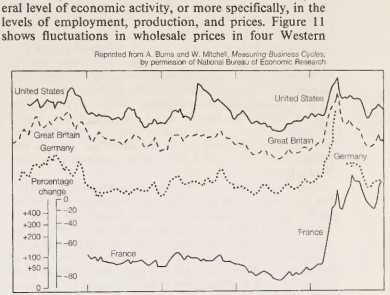


Figure 11: Wholesale price indexes for United States, Great Britain, Germany, and France, 1790–1940.

industrialized countries over the period from 1790 to 1940. Though some regularities in price movements are apparent, it is possible to ask whether the movements are regular enough to be called cycles.

The word *cycle* derives from the Greek word for circle. An object moving around a circle returns to its starting point; a wave motion, with upward and downward curves, may also be considered a cycle. The various movements characteristic of economic activity are not always as regular as waves, and for this reason some prefer to call them fluctuations.

The varieties of cycles

There are many types of economic fluctuations. Because of the complexity of economic phenomena, it may be that there are as many types of fluctuations or cycles as there are economic variables. There are daily cycles in commuter traffic or the consumption of electricity, to cite only two examples. Almost every aspect of economic life displays seasonal variations: sales of coal or ice, deposits in savings banks, monetary circulation, agricultural production, purchases of clothing, travel, housing, entertainment, and so on. As one lengthens the span of observation, one finds new kinds of fluctuations such as the hog cycle and

the wheat cycle, the inventory cycle, and the construction cycle. Finally, there are expansions or contractions of general economic activity that extend over periods of years.

Historical studies of cycles. Modern economic history has recorded a number of periods of difficult times, often called depressions, during which the business economy was marked by sudden stock market declines, commercial bankruptcies, bank failures, and mounting unemployment. Such crises were once looked upon as pathological incidents or catastrophes in economic life, rather than as a normal part of it. The notion of a "cycle" implies a different view. The following examples represent some of the attempts theorists have made to explain and predict business cycles.

Juglar's eight-year cycle. The first authority to explore economic cycles as periodically recurring phenomena was probably Clément Juglar, a French physician, in 1860. Other writers who developed Juglar's approach suggested

The wealth of data

that the cycles recur every nine or 10 years and distinguished three phases, or periods, of a typical cycle: prosperity, crisis, and liquidation. Subsequent analysis has tended to designate 1825, 1836, 1847, 1857, 1866, 1873, 1882, 1890, 1900, 1907, 1913, 1920, and 1929 as initial years of crisis. If that is correct, then the average interval between them was eight years, rather than nine or 10 as suggested by Juglar. In the years since 1929, the regularity of business fluctuations has been somewhat offset by government anticyclical policies and by increases in complexity and specialization that have reduced dependence on any one economic sector.

The so-called Juglar cycle has often been regarded as the true, or major, economic cycle, but several smaller cycles have also been distinguished. Close study of the interval between the peaks of the Juglar cycle suggests that partial setbacks occur during the expansion, or upswing, and that there are partial recoveries during the contraction, or downswing. These smaller cycles generally coincide with changes in business inventories, lasting an average of 40 months. According to this analysis, other small cycles were said to result from changes in the demand for and supply of particular agricultural products such as hogs (three to four years), cotton (two years), and beef (five years in The Netherlands). Hide and leather production was thought to fluctuate in an 18-month cycle.

Kondratev's waves. Longer cycles have also been studied. The construction industry has been found to have cycles of 17 to 18 years in the United States and 20 to 22 years in England. Finally, there are the long waves, or so-called Kondratev cycles, named for a Russian economist, Nikolai D. Kondratev, who showed that in the major Western countries during the 150 years from 1790 to 1940 it was possible to distinguish three periods of slow expansions and contractions of economic activity averaging 50 years in length:

1. 1792–1850	Expansion: 1792–1815	23 years
	Contraction: 1815–50	35 years
2. 1850–96	Expansion: 1850–73	23 years
	Contraction: 1873–96	23 years
3. 1896–1940	Expansion: 1896–1920	24 years
	Contraction: 1920–40	20 years

Only these three Kondratev waves have been observed.

Some students of business cycles have analyzed them by statistical methods, in the hope of finding regularities that are not immediately apparent. One speculative theory has held that the larger cycles were built up from smaller ones. Thus, two seasonal cycles would produce a two-year cycle, two of which would produce a four-year cycle; two four-year cycles would become an eight-year, or Juglar, cycle, and so on. The hypothesis is not widely accepted.

Patterns of depressions and upswings. Cycles of varying lengths are closely bound up with economic growth. In

Cycles in history

19th-century Germany, for example, upswings in total economic activity were associated with the growth of the railroad, metallurgy, textile, and building industries. Periodic crises brought slowdowns in growth. The crisis of 1873 led to a wave of financial and industrial bankruptcies; recovery started in 1877, when iron production ceased to fall, and by 1880 a new upswing was under way. The recession of 1882 was less severe than the previous one, but a slump that began in 1890 led to a serious depression, with complaints of overproduction.

The year 1890 was also one of financial crisis in England and the United States. The British banking house of Baring Brothers failed, partly because of a revolution in Argentina. English pig-iron production fell from 8,300,000 tons in 1889 to 6,700,000 in 1892, and unemployment increased. That depression might have been less severe but for the international financial crisis, especially intense in the United States, where in 1893 a stock market panic led to widespread bank failures.

The recession of 1900 was followed by an unusually vigorous upsurge in almost all of the Western economies. U.S. pig-iron production increased by more than 150 percent during the expansion, which lasted until 1907; building permits more than doubled; and freight traffic rose by more than 50 percent. Prices rose more and more rapidly as the U.S. economy approached full employment.

Deviations from cycle patterns. Cycles are compounded of many elements. Historical fluctuations in economic activity cannot be explained entirely in terms of combinations of cycles and subcycles; there is always some factor left over, some element that does not fit the pattern of other fluctuations. It is possible, for example, to analyze a particular fluctuation into three principal components: a long component or trend; a very short, seasonal component; and an intermediate component, or Juglar cycle. But these components cannot be found exactly recombined in another fluctuation because of a residual element in the original fluctuation that does not have a cyclical form. If the residual is small, it might be attributed to errors of calculation or of measurement. Or, the residual might be regarded as the result of such accidental events as epidemics, floods, earthquakes, riots, strikes, revolutions, or wars, which obviously cannot be fitted into a recurring pattern. On a more sophisticated statistical level, it can be treated as "random movement." If the random element is always present, it becomes an essential element of the analysis to be dealt with in terms of probability.

For practical purposes, it would be useful to know the typical shape of a cycle and how to recognize its peak and trough. A great amount of work has been done in what may be called the morphology of cycles. In the United States, Arthur F. Burns and Wesley C. Mitchell have based such studies on the assumption that at any specific time there are as many cycles as there are forms of economic activity or variables to be studied and have tried to measure these in relation to a "reference cycle," which they artificially constructed as a standard of comparison. The object in such studies was to describe the shape of each specific cycle, to analyze its phases, to measure its duration and velocity, and to measure the amplitude or size of the cycle.

In studying various cycles, it has been possible to construct "lead and lag indicators"—that is, statistical series with cyclical turning points consistently leading or lagging behind the turns in general business activity. Researchers using these methods have identified a number of series, each of which reaches its turning point from two to 10 months before the turns in general business activity, and another group of series, which has followed the turns in business by two to seven months. Examples of leading series include published data for new business orders, residential building contracts, the stock market index, business failures, and the length of the average workweek. These and other leading indicators are widely used in economic forecasting.

Theories of economic fluctuation. A satisfactory explanation of cycles must isolate the forces and relationships that tend to produce these recurrent movements. There have been many theories of the business cycle. An understanding of them requires analysis of some of the factors that can cause cyclical movements. Theories of the business cycle, or, more properly, of economic fluctuations, may be classified in two groups: those that ascribe cyclical movements to external forces (exogenous factors) and those that attribute the fluctuations to internal forces (endogenous factors).

Keynesian theories. Some theorists examine the relationship between investment and consumption. According to Keynesian economists, any new expenditure—*e.g.*, on building a road or a factory—generates several times as much income as the expenditure itself. This occurs because those who are paid to build the road or factory will spend more of what they receive; their expenditures will thus become income for others, who will in their turn spend most of what they receive. Every new act of investment will, thus, have a stimulating effect on aggregate income. This relationship is known as the *investment multiplier.* Of itself, it cannot produce cyclical movements in the economy; it merely provides a positive impulse in an upward direction.

To the relationship between investment and consumption must be added that between consumer demand and investment. An increase in demand for refrigerators, for example, will eventually require increased investment in the facilities for producing them. This relationship is known as the *accelerator;* and it implies that an increase in national income will stimulate investment. As with the multiplier, it cannot of itself explain cyclical movements; it merely accounts for a fundamental instability.

It can be shown, however, that the multiplier and accelerator in combination may produce very strong cyclical movements. Thus, when an increase in investment occurs, it raises income by some larger amount, depending on the value of the multiplier. That increase in income may in turn induce a further increase in investment. The new investment will stimulate a further multiplier process, producing additional income and investment. In theory, the interaction might continue until a point is reached at which such resources as labour and capital are being fully utilized. At that point—with no increase in employment and, therefore, no rise in consumer demand—the operation of the accelerator would cease. That halt in demand, plus the lack of new capital, would cause new investment to decline and workers to be laid off. The process thus would go into reverse. The fluctuations in national income could take various forms, depending on the characteristics of the economy and the way in which the population allocated its income between consumption and savings. Such spending habits, of course, affect both the levels of consumer demand and capital investment. This theoretical analysis does not explain actual economic fluctuations; it is merely an aid to understanding them.

The analysis can be made more realistic by taking into account three other factors. First, since the theoretical, wide-swinging cycles engendered by the interaction of the multiplier and the accelerator are observed to occur only within narrow limits, one may assume that although the economy has an inherent tendency to swing very widely there are limits beyond which it cannot go. The upper limit to the swings would be the point at which full employment or full capacity is reached; the lower limit is more difficult to define, but it would be established when the forces that make for long-term economic growth begin to operate. Thus, the upswing of a cycle stops when it meets the upper limit; and the downswing stops at the lower limit, resulting in continuous cyclical movements with an overall upward trend—a pattern corresponding to the one found in history.

Upper and lower limits to swings

In economic life, there are many time lags: between the decision to invest and the completion of the project; between the farmer's decision to raise hogs and the arrival of pork chops at the store; between prices at the time of a decision and prices at the time the action is completed.

Random shocks, or what economists call exogenous factors, constitute the third type of phenomena affecting business cycles. These are such external disturbances to the system as weather changes, unexpected discoveries, political changes, wars, and so on. It is possible for such external impulses to cause cyclical motions within the system, in much the same way that striking a rocking horse with a stick will cause the horse to rock back and forth. The length of the cycle will be determined by the internal relationships of the system, but its intensity is governed by the external impulse.

Agricultural and climatic theories. Perhaps the oldest theories of the business cycle are those that link their cause to fluctuations of the harvest. Since crops depend upon natural factors that in turn may be affected by biological or meteorological cycles, such cycles will transmit their effects through the harvests to the rest of the economy. The 19th-century English economist William Stanley Jevons thought he had found the key to such a process in the behaviour of sunspots, which seemed to display a 10-year cycle. His naive explanation could not long withstand critical examination. It had a certain interest, however, in suggesting a causal factor that was completely detached from the economic system and one that could not be influenced by it in turn. Agricultural theories made sense at a time when agricultural goods represented 40 to 60 percent of the output of an advanced economy. By the turn of the 21st century, however, agriculture's contribution to an advanced economy's output might be 5 percent or less.

Psychological theories. A number of writers have explored mass psychology and its consequences for economic behaviour. Individuals are strongly influenced by the beliefs of the group or groups to which they belong. There are times when the general mood is optimistic, and others when it is pessimistic. An English economist, Arthur C. Pigou, in his *Industrial Fluctuations* (1927), put forward a theory of "noncompensated errors." He pointed out that if individuals behave in a completely autonomous way their errors in expectations will tend to offset each other. But if they imitate each other, their errors will accumulate until they acquire a global magnitude that may have powerful economic effects. This follow-the-crowd tendency obviously operates as a factor in the ups and downs of the stock exchanges, financial booms and crashes, and the behaviour of investors. One can say, however, that the psychological factor is not enough to explain economic fluctuations; moods of optimism and pessimism must themselves rest upon economic factors.

Political theories. Some observers have maintained that economic fluctuations result from political events. It is obvious that such events as the Napoleonic Wars, World Wars I and II, and even the Korean War of 1950–53 have had strong economic consequences. Even the imposition of a tax or an import restriction may have some dynamic effect upon the economy. In the United States, some economists have speculated that incumbent political leaders pressure the chairman of the Federal Reserve System to loosen monetary policy in advance of an election as a means of fostering prosperity. The question is whether such political factors are capable of producing cyclical movements.

Effects of wars and taxes

Technological theories. Ever since the Industrial Revolution at the end of the 18th century, technical innovations have followed each other without end but not without pause. There have been periods of innovation and quieter periods in which the innovations were being absorbed. More recently, computers and the Internet have influenced almost all aspects of economic production. It is possible that, if a rhythm could be found in these waves of change, the same rhythm might be responsible for corresponding movements in the economy. But it is equally possible that the technical innovations themselves have been dictated by the prior needs of the economy.

Demographic theories. Even population has been postulated as a cause of economic fluctuations. There are, undeniably, cyclical movements of population; it is possible to find fluctuations in the rates of marriage, birth, mortality, and migration; but the extent to which such fluctuations may have been caused by economic conditions is not clear.

Monetary theories. Some writers have ascribed economic fluctuations to the existence of money. Changes in the money supply do not always conform to underlying economic changes, and it is not difficult to see how this lack of coordination could produce disturbances in the economic system. Thus, an increase in the total quantity of money, if it is not matched by an increase in economic activity, will tend to produce higher prices; the higher prices may stimulate an investment boom, and so on.

The banking system, with its ability to expand the supply

of credit in a time of boom and to contract the supply of credit in time of recession, may in this way amplify small economic fluctuations into major cycles of prosperity and depression. Some theorists, such as Knut Wicksell, have emphasized the influence of the rate of interest: if the rate fixed by the banking system does not correspond to the "natural" rate dictated by the requirements of the economy, the disparity may of itself induce an expansion or contraction in economic activity.

Underconsumption theories. In a progressive economy, production tends to expand more rapidly than consumption. The disparity results from the unequal distribution of income; the rich do not consume all their income, while the poor do not have sufficient income to meet their consumption needs. This imbalance between output and sales has led to theories that the business cycle is caused by overproduction or underconsumption. But the basic, underlying cause is society's inadequate provision for an even flow of savings out of the excess of what is produced over what is consumed. In other words, saving is out of step with the requirements of the economy; it is improperly distributed over time.

Imbalance between output and sales

Investment theories. The fact that changes in the supply of savings, or loanable funds, are not closely coordinated with changes in the rest of the economy lies at the heart of the numerous theories that link investment imbalance to the business cycle. Savings accumulate when there is no immediate outlet for them in the form of new investment opportunities. When times become more favourable, these savings are invested in new industrial projects, and a wave of investment occurs that sweeps the rest of the economy along with it. It is in this context that the tools of analysis—the accelerator and the multiplier—find their application: the new investment creates new income, which in turn acts as a further stimulus to investment. An early observer of this phenomenon, a Russian economist, Mikhayl Tugan-Baranovsky, in 1894 published a study of industrial crises in England in which he maintained that the cycle of investment continues until all the capital funds have been used up. Bank credit expands as the cycle progresses. Disproportions then begin to develop among the various branches of production as well as between production in general and consumption. These imbalances lead to a new period of stagnation and depression.

Rational expectations and "real" business cycles. In the early 1970s, American economist Robert Lucas laid out what has come to be known as the "Lucas critique" of both Keynesian and monetarist theories of the business cycle. Such theories, he argued, ignored the effects of rational economic behaviour. Lucas observed that people do act rationally, for example, when they try to anticipate the consequences of a change in government policy. In some cases, their actions will offset the effects the government had hoped to achieve through the policy change.

Lucas won the 1995 Nobel Prize in Economic Science for development of the rational expectations theory. His work spurred other 20th-century economists who posited that business fluctuations are not due to changes in monetary policy; instead, these fluctuations stem from underlying changes in the economy. This came to be called real business cycle theory. In this view, economic fluctuations are not bad at all but, instead, are a healthy adjustment to underlying conditions—an adjustment that is necessary if economic growth is to continue. (H.Gu./Ed.)

STABILIZATION THEORIES AND POLICIES

The ultimate objective of research into the problems of economic instability (including fluctuations in output, employment, and prices) is to provide the foundation for stabilization policy—that is, for the systematic use of fiscal and monetary policies to improve an economy's performance. The main tasks, therefore, are to explain how levels of prices, output, and employment are determined and, on a more applied level, to furnish predictions of changes in these variables—predictions on which stabilization policy can be based.

Keynesian analysis. The problems of economic stability and instability have, naturally, been of concern to economists for a very long time. But, as a special field of investigation, it emerged most strongly from the confluence of two developments of the depression decade of the 1930s. One was the development of national income statistics; the other was the reorientation of theoretical thinking often referred to as the "Keynesian revolution."

To understand why the theoretical contributions of Keynes were regarded as so important through much of the 20th century, one must examine the workings of a modern economy. Such an economy comprises millions of people engaged in millions of distinct activities; these activities include the production, distribution, and consumption of all of the different goods and services that a modern economy provides. Some of the economic units are large, with hierarchies of executives and other managerial specialists who coordinate the productive activities of thousands or tens of thousands of people. Aside from these relatively small islands of preplanned and coordinated activity, most of the population pursues its myriad economic tasks without any overall supervised direction. It resembles an immensely complicated, continuously changing puzzle that is continually being solved and solved again through the market system. A breakdown in the coordination of activities, such as occurred in the depression decade of the 1930s, is very rare—in fact, it happened on that scale only once—or this system of organization would not survive. The way in which the economic puzzle is solved without anyone thinking about it has been the broad main theme of economic theory since the time of the English economist Adam Smith (1723–90).

The problem of coordination. If one singles out a particular household from the millions of economic units and studies it over a period of time, one can draw up a budget of that household's transactions. The budget will come out as a long list of amounts sold and amounts bought. If at any time this economic unit had tried to do something different from what it actually did (cutting down, say, on meat purchases to buy another pair of shoes), the solution of the economic puzzle would have been correspondingly different. At the prevailing prices the supply of meat would have exceeded the demand, and the demand for shoes would have exceeded the supply.

The price system; role of incentives

The point Keynes made, right or wrong, was that, if the economy were to function as a coordinated system, the activities of each economic unit must be somehow controlled—and controlled quite precisely. This is done through price incentives. By raising the price of a good (relative to the prices of everything else), any economic unit can, generally speaking, be made to demand less of it or to supply more of it; by lowering the price, it can be made to demand more or to supply less. Through the confluence of prices, an individual unit is thus led to fit its activities into the overall puzzle of market demands and supplies. If economic units could not be controlled in this fashion, the market-organized system could not possibly function.

Keynesians therefore believe that in any given situation there exists, theoretically, one and only one list of prices that will make the puzzle come out exactly right. But the amounts that economic units choose to supply or demand of various goods at any given price list depend on numerous factors, all of which change over time: the size of the population and labour force; the stock of material resources, technology, and labour skills; "tastes" for particular consumer goods; and attitudes toward consumption as against saving, toward leisure as against work, and so on. Government policies—tax rates, expenditures, welfare policies, money supply, the debt—also belong among the determinants of demand and supply. A change in any of these determinants will mean that the list of prices that previously would have equilibrated all of the different markets must be changed accordingly. If prices are "rigid," the system cannot adjust and coordination will break down.

Price flexibility. For coordination of activities to be preserved (or restored) when the economy is disturbed by changes in these determinants, something still more is required: each separate price must move in a direction that will restore equilibrium. This necessity for prices to adjust in certain directions may be expressed as a com-

munications requirement. To put it in somewhat extreme form: for a given economic unit to plan its activities so that they will "mesh" with those of others, it must have information about the intentions of everyone else in the system. When one of the determinants underlying market supplies and demands changes so as to disequilibrate the system, ensuing price movements must communicate the requisite information to everyone concerned.

One may suppose, for example, that in some period of political crisis the supply of crude oil from the Middle East is cut off. The immediate result will be a worldwide excess demand for oil and oil products of large proportions—that is, supply will fall far short of demand at going prices. At the same time, those who derive their income from Middle East oil production will have their incomes reduced, and excess supplies will emerge in the markets for the goods on which those incomes previously were spent. For the system to adjust, orders will have to go out to all demanders to cut down on their consumption of oil and for all other suppliers of oil to increase their output so that the gap between demand and supply can be closed. This is, in effect, what a rise in the world price of oil and oil products will accomplish—millions of gasoline and heating oil users the world over will respond to the pinch of higher prices, and the higher prices will also create a profit incentive for supply to be increased. (Falling prices will, in an analogous manner, close the gaps in the markets in which the initial disturbance caused excess supplies to develop.)

Prices that are not rigid for some institutional reason will move in response to excess demands and excess supplies. When demand exceeds supply, disappointed buyers will bid up the price; when supply exceeds demand, unsuccessful suppliers will bid it down. This mechanism solved the excess demand for the oil problem in the illustration above. The question, however, is whether throughout the system as a whole it will always act so as to move each of the prices toward its general equilibrium value.

Keynes said no. He maintained that there can be conditions under which excess demands (or supplies) will not be "effectively" communicated so that, although certain prices are at disequilibrium levels, no process of bidding them away from these inappropriate levels will get started. This is the flaw in the traditional conception of the operation of the price system that prompted Keynes to introduce the concept of "effective demand." To pre-Keynesian economists the implied distinction between "effective" and (presumably) "ineffective" demand would have had no analytical meaning. The logic of traditional economic theory suggested two possibilities that might make the price system inoperative: (1) that, in some markets, neither demanders nor suppliers respond to price incentives, so that a "gap" between demand and supply cannot be closed by price adjustments and (2) that, for various institutional reasons, prices in some markets are "rigid" and will not budge in response to the competitive pressures of excess demands or excess supplies. Keynes discovered a third possibility that, he argued, was responsible for the depth and duration of severe depressions: under certain conditions, some prices may show no tendency to change even though desires to buy and to sell do not coincide in the respective markets and even though no institutional reasons exist for the prices to be rigid.

Causes of depression

Say's Law. Many writers before Keynes raised the question of whether a capitalist economic system, relying as it did on the profit incentive to keep production going and maintain employment, was not in danger of running into depressed states from which the automatic workings of the price mechanism could not extricate it. But they tended to formulate the question in ways that allowed traditional economics to provide a demonstrable, reassuring answer. The answer is known in the economic literature as Say's Law of Markets, after the early 19th-century French economist Jean-Baptiste Say.

For western Europe, the 19th century was a period of rapid economic growth interrupted by several sharp and deep depressions. The growth was made possible in large measure by new modes of organizing production and new technologies, such as the spreading use of steam power. Was it possible that output might grow so great that there would not be a market for it all? Say's Law denied the possibility. "Supply creates its own demand," ran the answer. More precisely, the law asserted that the sum of all excess supplies, evaluated at market prices, must be identically equal to the sum of the market values of all excess demands. It could be neither more nor less. In the theoretical system of traditional economics, any inequality between these sums would quickly work itself out.

Excess demand for money

An important special case should be noted. The good in excess demand might, for instance, be money. One possibility, then, is excess supply for all the other goods, matched by an excess demand for money. A situation with excess demand for money matched by an excess supply of everything else is one in which the level of all money prices is too high relative to the existing stock of money. If this is the only trouble, however, Say's Law suggests a relatively simple remedy: increase the money supply to whatever extent required to eliminate the excess demand. The alternative is to wait for the deflation to work itself out. As the general level of prices declines, the "real" value of the money stock increases; this too, will, in the end, eliminate the excess demand for money.

Involuntary unemployment. Another possible cause of a general depression was suggested by Keynes. It may be approached in a highly simplified way by lumping all occupations together into one labour market and all goods and services together into a single commodity market. The aggregative system would thus include simply three goods: labour, commodities, and money. The Table provides a rough outline (a full treatment would be both technical and lengthy) of the development of a "Keynesian" depression.

Model of a "Keynesian" Depression

	Excess Demand (ED) or Excess Supply (ES) for:			notes
	labour	commodities	money	
Initial state	0	0	0	Equilibrium
State 2	0	ES_C	ED_M	$pES_C = ED_M$
State 3	ES_L	0	ED_M	$wES_L = ED_M$
State 4	ES_L	ED_C	0	$wES_L = ED_C$

One may begin by assuming (line 1) that the system is in full employment equilibrium—that is, prices and wages are at their equilibrium levels and there is no excess demand. Next the model may be put on the path to disaster by postulating either (1) some disturbance causing a shift of demand away from commodities and into money or (2) a reduction in the money supply. Either event will result in the situation described in the Table as State 2, but the one assumed is a reduction in the money supply by, say, 10 percent. The result is shown in the right-hand column of the Table, where the quantity of commodities supplied minus the quantity demanded multiplied by the price level (p) is equal in value to the excess demand for money.

If money wages and money prices could immediately be reduced in the same proportion (10 percent), output and employment could be maintained, and profits and wages would be unchanged in "real" terms. If money wages are initially inflexible, however, business firms cannot be induced to lower prices by 10 percent and maintain output. In this example they maintain prices in the neighbourhood of the initial price level—prices, then, are also "inflexible"—and deal with the excess supply by cutting back output and laying off workers. Reducing supply eliminates the excess supply of commodities by throwing the burden of excess supply back on the labour market. Thus, output and employment (which are "quantities") give way before prices do. This brings us to State 3 where, as in the Table, the excess supply of labour times the money wage rate (w) equals the excess demand for money in value.

The lack of effective demand

If, with the system in this state, money wages do not give way and the money supply is not increased, the economy will remain at this level of unemployment indefinitely. One should recall that the only explanation for persistent unemployment that the pre-Keynesian economics had to offer was that money wages were "too high" relative to the money stock and tended to remain rigid at that level.

Money wages might, nevertheless, give way so that, gradually, both wages and prices go down by 10 percent—that is to say, a reduction of the size that would have solved the entire problem had it occurred immediately (*before* unemployment could develop). This is shown in the last line of the Table, which represents (albeit crudely) what Keynes described as a state of "involuntary unemployment" and explained in terms of a failure of "effective demand."

In State 4, it is assumed, the excess demand for money is zero. Hence there is, at least temporarily, no tendency for money income either to fall further or to rise. The prevailing level of money income is too low to provide full employment. The excess supply of labour and the corresponding excess demand for commodities (of the same market value) show State 4 to be a disequilibrium state. The question is why the state tends to persist. Why is there no tendency for income and output to increase and to absorb the unemployment? Specifically, why does not the excess demand for commodities induce this expansion of output and absorption of unemployment?

Basically, the answer is that the unemployed do not have the cash (or the credit) to make the excess demand for commodities *effective*. The traditional economic theory would postulate that, when actual output is kept at a level below that of demand, competition between unsuccessful potential buyers would tend to raise prices, thereby stimulating an expansion. But this does not occur. The unemployed lack the means to engage in such bidding for the limited volume of output. The excess demand for commodities is not effective. It fails to produce the market signals that would induce adjustments of activities in the right direction. Business firms, on their side of the market, remain unwilling to hire from the pool of unemployed—even at low wages—because there is nothing to indicate that the resulting increment of output can actually be sold at remunerative prices.

Keynes called this "involuntary unemployment." It was not a happy choice of phrase since the term is neither self-explanatory nor very descriptive. Some earlier analysts of the unemployment problem had, however, tended to stress the kind of deadlock that might develop if workers held out for wages exceeding the market value of the product attributable to labour or if business firms insisted on trying to "exploit" labour by refusing to pay a wage corresponding to the value of labour's product. With the term "involuntary unemployment," Keynes wanted to emphasize that a thoroughly intractable unemployment situation could develop for which neither party was to blame in this sense. His theory envisaged a situation in which both parties were willing to cooperate, yet failed to get together. An effective demand failure might be described as "a failure to communicate."

The failure of the market system to communicate the necessary information arises because, in modern economies, money is the only means of payment. In offering their labour services, the unemployed will not demand payment in the form of the products of the individual firms. If they did, the excess demand for products would be effectively communicated to producers. The worker must have cash in order to exercise effective demand for goods. But to obtain the cash he must first succeed in selling his services.

When business begins to contract, the first manifestation is a decrease in investment that causes unemployment in the capital goods industries; the unemployed are deprived of the cash wage receipts required to make their consumption demands effective. Unemployment then spreads to consumer goods industries. In expansion, the opposite occurs: an increase in investment (or in government spending) leads to rehiring of workers out of the pool of unemployed. Re-employed workers will have the cash with which to exert effective demand. Hence business will pick up also in the consumer goods industries. Thus the theory suggests the use of fiscal policy (an increase in government spending or a decrease in taxes) to bring the economy out of an unemployment state that is due to a failure of effective demand.

Another observation may be made on Keynes's doctrine of effective demand. The fact that the persistence of unemployment will put pressure on wages also turns out to be a problem. The assumption in the foregoing discussion was that money wages were at the equilibrium level. Unemployment will tend to drive them down. Prices will tend to follow wages down, since declining money earnings for the employed will mean a declining volume of expenditures. In short, both wages and prices will tend to move away from, rather than toward, their "correct" equilibrium values. Once the economy has fallen into such a situation, Keynes pointed out, wage rigidity may actually be a blessing—a paradoxical conclusion from the standpoint of traditional economics.

National income accounting. *The circular flow of income and expenditure.* A proper understanding of income and expenditure theory requires some acquaintance with the concepts used in national income accounting. These accounts provide quantitative data on national income and national product. Reliable information on these was, for the most part, not available to economists working on problems of economic instability before the 1930s. Modern economics differs from earlier work most markedly in its quantitative, empirical orientation. The development of national income accounting made this possible.

The definitions of the major components of national income and product may, accordingly, be introduced in the course of explaining income and employment theory. The basic characteristic of the national income accounts is that they measure the level of economic activity in terms of both product supplied and income generated. Correspondingly, national income analysis divides the economic system into distinct *sectors*. The simplest approach uses two sectors: a business sector and a household sector. All product is regarded as created by the business sector (thus, self-employed persons have to be treated as businesses in earning their income and as households in disposing of it). Final goods output is divided into two components: consumer goods produced for sale to households and investment goods for sale to firms. Similarly, all income is generated in the business sector and none of it in the household sector (nonmarket activities, such as the work of homemakers or home improvements, are not counted in national product and income). The level of income generated equals the market value of final goods output.

Economic sectors

Next is the household sector. All resources in the economy ultimately belong to households. The households, therefore, have claim to all of the income generated through the utilization of these resources by firms in creating the national product. Not all of the income is, however, actually paid out to households, since corporations retain part of their earnings. In building a simple model of the economy, one can disregard the "gross business saving" item of the national income accounts and deal with income as if it were all paid out (which means adopting the fiction that retained earnings are first paid out to shareholders who then reinvest the same amount in the same firms). The households, finally, dispose of their income in two ways: as expenditure on consumption goods and as saving.

The foregoing discussion has made two accounting statements involving income. First, income *generated* (Y) equals the value of consumption goods output (C^y) plus the value of investment goods output (I): $Y \equiv C^y + I$. Second, consumption goods expenditures (C^e) plus savings (S) equal income *disposal*: $Y \equiv C^e + S$. Both equalities hold simply because of the way that the variables are defined in the national income accounts. They hold true, moreover, whatever the actual level of income happens to be. Such equalities, which are true simply by definition, are called *identities* (and are marked as such by using the sign $\equiv$ instead of the usual equality sign). Another accounting convention may be noted here. Investment (I) is defined to include any discrepancy between consumer goods produced and consumer goods sold. If production exceeds sales, the unsold goods are part of inventory investment; if sales exceed output, inventory investment is negative, and I is reduced by the corresponding amount. It follows that C^y and C^e must be identically equal, so that it becomes unnecessary to distinguish between them by superscript. Since income generated is identically equal to income disposal, finally, it is clear that actual investment must always equal actual saving: $I \equiv S$. Investment is the value of additions to the system's stock of capital. Saving is the increase in the value of the household sector's wealth. For the system as a whole, the two must be equal.

Figure 12 shows the circular flow of income and expenditures connecting the two sectors. Investment and consumption expenditures add up to the *aggregate demand* for final goods output. The value of final goods output is paid out by the business sector as income to the household sector. The major part of income goes back to the business sector as expenditures on consumption goods; the remainder is allocated by households to saving. Corresponding to the counterclockwise money flow (but not shown) is the clockwise flow of the things that the money is paid for: labour and other resource services from households to firms in exchange for money income; consumer goods and services in exchange for consumption expenditures from firms to households; and equities, bonds, and other debt instruments issued by firms in return for the funds saved by households.

Figure 12 shows a break in the flow of saving as it passes

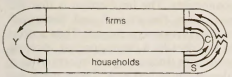


Figure 12: The circular flow of income expenditures (see text).

into investment. From the accounting standpoint—where investment necessarily equals saving—there is no rationale for this. It has been done here to focus attention on the point in the circular flow that, in the income–expenditure theory, represents the causal nexus in the income-determining process. This theory, in its simplest form, is the next topic.

A simple income–expenditure model. Because accounting identities—between gross national product and gross national income, between saving and investment, and so on—express relationships that must hold whatever the level of income, they cannot be used to explain what determines the particular level of income in a given period or what causes the level of income to change from one period to the next. The explanation of what happens must be based on statements about the behaviour of the participants in the economic system; in the present context, this means the behaviour of firms and households.

The dynamics of income and expenditure

The following oversimplified model of an economy assumes that the business sector will be satisfied to maintain *any* given level of output as long as *aggregate demand* (that is, expenditures on final goods) exactly equals the volume of income generated at that level of output. If, in a given period, aggregate demand exceeds the income payments made by firms in producing that period's output, firms will be expanding in the next period; if aggregate demand falls short of the income payments made, firms will contract in the next period. The naïveté of this supply hypothesis is evident from the fact that the behaviour of firms is described without any reference to the costs of their inputs or to the price of their outputs; the business sector passively adapts output and income generated to the level of aggregate demand. In this model, the level of income is entirely determined by aggregate demand. Firms will act so as to maintain that income flow if, and only if, the exact same amount that they pay out as incomes "comes back to them" in the form of spending on final goods output. If aggregate demand shrinks, production and employment will decline and there will be downward pressure on the price level; if aggregate demand swells, there will be an inflationary problem.

Leakages and injections

In the system of Figure 12, all of the income generated accrues to households. Households allocate their income to consumption and saving. With consumption there is no problem—it constitutes spending on final goods. Saving, however, does not constitute spending on final goods output. This part of the income generated by the business sector does not automatically come back to it in the form of revenue from sales. Saving, therefore, may be treated as a *leakage* from the circular flow.

Investment, which consists of spending of capital by the business sector on new plant and equipment and on desired additions to inventories, is, in the same terminology, an *injection* into the circular flow. If, for example, investment and saving each amount to $20,000,000 per year, the leakage and the injection will balance. But if saving is $20,000,000 per year and the injection of investment expenditures is only $10,000,000 per year, there will be a disequilibrium. Unsold goods will accumulate at an annual rate of $10,000,000. The business sector, however, will not rest content with this state of affairs but will act to reduce output, employment, and (perhaps) prices. Households will be forced to reduce their consumption spending. The reduction of income will go on until the planned (or desired) rates of saving and investment become equal. A similar argument will show that, if the leakage of planned saving were to fall short of the injection of planned investment, the level of income would rise.

When income is at a level such that there is no ongoing tendency for it to change in either direction, the system is in "income equilibrium." The simple system depicted in Figure 12 is in income equilibrium when the condition shown by this equation is fulfilled: $I = S$. This is not, however, the accounting identity discussed earlier. The symbols I and S now refer to planned, or desired, magnitudes, which may very well be unequal. When planned investment exceeds planned saving, income will be rising. When planned saving exceeds planned investment, income will be falling. An equivalent way of stating the above "equilibrium condition" is to write $Y = C + I$. In this equation the left-hand side is actual income and the right-hand side is planned aggregate demand.

This is the simplest class of income-determination model. It makes no allowance for international trade or government economic activity. Those may be treated in the same way that saving and investment were treated—as leakages or injections. Thus exports constitute spending by foreign nationals on domestic goods—an injection. Imports constitute spending out of domestic income on foreign goods—a leakage. Taxes are taken out of the circular flow—a leakage—whereas government expenditures are an injection. The effects of these leakages and injections on the level of income are analogous to those of saving and investment. If income is initially at an equilibrium level, an increase in a leakage (if not at the same time offset by a decrease in another leakage or an increase in an injection) will cause income to fall. An increase in an injection (not offset by a decrease in another injection or an increase in a leakage) will cause income to rise. An income equilibrium is reached when the sum of all leakages is balanced by the sum of all injections.

The multiplier

The simple income–expenditure model of the economy is not a complete model. It suffices to show only the *direction* of the change in income that would result from, say, a decline in planned investment (or a rise in taxes or a decline of exports). It does not show the *extent* of the income change.

To do this the model must be expanded to include a description of how consumers spend their incomes. For the sake of the exposition, one may assume that the spending of households varies according to the size of their incomes. A simple way of putting this is the following equation: $C = a + by$. In this equation the coefficient a is a constant indicating the amount that households will spend on consumption independently of the level of income received in the current period, and the coefficient b gives the fraction of each dollar of income that will be spent on consumption goods.

If one were able to obtain reliable quantitative information on the volume of investment spending being planned and on the coefficients a and b of the "consumption function" above, one could then calculate the value of aggregate demand ($C + I$) for every possible level of income Y. Only one of these alternative levels of income is an equilibrium one; that is, one for which aggregate demand will ensure that all of the income paid out by firms "comes back" to the business sector as spending on final goods. The equilibrium condition is: $Y = C + I$.

Figure 13 shows how the level of income in the system is determined, on the assumption that investment is $20,000,000, that the coefficient a is $20,000,000, and that the coefficient b (the fraction of each dollar of income that consumers will spend) is 0.6. The horizontal axis measures income, the vertical, aggregate demand ($C + I$). The line drawn at a 45° angle (from 0) contains all of the points at

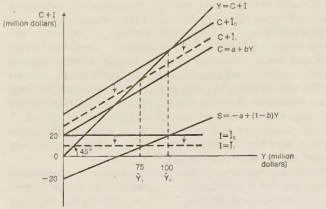


Figure 13: Relation between income and aggregate demand (see text).

which *suppliers* might be in equilibrium; *i.e.,* the points in the space at which aggregate demand would have the same value as income. The investment schedule (marked $I = \bar{I}_0$) is drawn parallel to the income axis at height 20, showing that investment spending does not depend on income. The consumption function (marked $C = a + by$) starts at 20 on the vertical axis (the value of a) and rises 60 cents for each dollar of income (the value of b) to the right. The aggregate demand schedule (marked $C + \bar{I}_0$) is obtained by the vertical summation of the C and $\bar{I}_0$ schedules. It contains all of the points at which *demanders* would be in equilibrium, showing, for each level of income, the volume of spending on final goods that they would be satisfied to maintain.

The only position that demanders and suppliers will both be satisfied to maintain is given by the intersection of the aggregate demand schedule with the 45° line. In Figure 13 this point ($\hat{Y}_0$) is found at an income level of $100,000,000. For this simple system, which has but one leakage and one injection, the equilibrium level of income may equally well be regarded as determined by the condition that planned saving equals planned investment. Since saving is defined as household income not spent on consumption (*i.e.*, $Y - C \equiv S$), one obtains (by substituting $a + by$ for c) the saving schedules $S = -a + (1 - b)\ Y$, which in Figure 13 is shown to intersect the investment schedule at $Y = $100,000,000.

Figure 13 shows what will happen if this equilibrium is disturbed. Consider a (temporary) situation in which income is running at *more than* $100,000,000 per year. At all levels of income to the right of $\hat{Y}_0$ aggregate demand $(C + \bar{I}_0)$ is seen to fall below supply as given by the 45° line. (Also, saving exceeds investment.) The business sector will not be willing to maintain this state of affairs but will contract. *An excess supply of final goods is associated with falling income.* Similarly, at income levels to the left of $\hat{Y}_0$, where investment exceeds saving, aggregate demand will exceed supply. *An excess demand for final goods is associated with rising income.*

Finally, Figure 13 shows how much income would fall as a result of a decline in investment by $10,000,000 per year (*cf.* the dotted lines). The decline in investment is shown by the shift of the investment schedule from $\bar{I}_0$ to $\bar{I}_1$, which results in a downward shift of the aggregate demand schedule from $C + \bar{I}_0$ to $C + \bar{I}_1$. The new income equilibrium ($\hat{Y}_1$) is found at $Y = $75,000,000.

Thus a change in investment spending (ΔI) of $10,000,000 is found to lead to a change in income (ΔY) of a larger amount, here $25,000,000, which is to say, by a multiple of 2.5. The reason is that, when the $10,000,000 is transmitted to households as income, households will increase their consumption spending by $6,000,000 ($b \times $10,000,000$). This rise in consumption spending again raises income, and of this additional income 60 percent is also spent on consumption—and so on. Each time, 40 percent of the increment to income "leaks" into saving. The relationship between the initial change in "autonomous spending" (ΔI) and the change in the level of income (ΔY), which will have taken place once this process has run its course, is given by:

$$\Delta Y = \left(\frac{1}{1-b}\right) \Delta I$$

where, following Keynes, the expression $\left(\frac{1}{1-b}\right)$ is called the "Multiplier."

Stabilization policy

The model of income determination presented above is exceedingly simple; it captures little of the complexity of a modern industrialized economy. It does, however, suggest one approach to the problem of stabilizing the economy at a high level of income and employment. Assuming that the consumption function is fairly stable (*i.e.,* that the level of consumption spending associated with any level of income can, with a fair degree of accuracy, be predicted on the basis of past experience), fluctuations in income may be attributed to changes in the other variables. Historical statistics show investment spending by private business to have been the most volatile of the major components of national income; changes in investment, therefore, tend (as in the example above) to be the focus of concern for one school of economists. The implication is that the government can manipulate "injections" and "leakages" so as to offset changes in private investment. Thus a drop in investment might be offset by a corresponding increase in government expenditures (increasing an injection) or a decrease in taxes (decreasing a leakage). These measures belong to fiscal policy.

Monetary policy. Another point of view holds that the fiscal approach presented above is misleading because it ignores the part played by monetary factors in determining the level of economic activity. The following discussion presents an alternative model, which, though equally simplistic, suggests that primary reliance be put on monetary policy.

"Money" in what follows may be taken to refer to currency (coins and notes) plus the checking deposit liabilities of commercial banks. For the sake of brevity, the model developed in the preceding section will be referred to as the income model. The naive quantity theory model that will be explained here may be labelled the money model.

The dynamics of income and money

The income model dealt with changes in money income in terms of the demand for and supply of output. The money model focusses on the supply of and demand for money. The income model explained the determination of the level of income in terms of relationships between its component flows. The money model emphasizes the relationship between money supply and income. The structure of the income model was based on the distinction between household and business (and government) sectors. In the money model, the distinction is between the banking sector (supplying the money) and the nonbanking sectors (the demanders). The concept of income is the same in both models.

In the money model, the supply of money is treated with the same simplicity that was accorded investment in the income model—as "autonomously" determined, which is to say that it is not affected by other factors: $M^s = \bar{M}$. This assumes that the central bank is able completely to control the stock of money, which is held at whatever level the bank desires.

The dynamic relationship in the income model was the consumption function. Here it is the money demand function. The amount of money demanded is assumed to vary with income (and, in this naive version of quantity theory, with nothing else). The simplest relationship between income and the demand for money would be: $M^d = kY$. Here, k is a constant. Since Y is a flow (measured per year) and M^d a stock (the average stock of money over the year), k has the dimension of a "storage period." If $k = 1/4$, for example, the equation states that the nonbanking public desires on the average to hold a cash balance that is equal to the total of three months' income.

Since there is a determined amount of money in the system, it can be in equilibrium only when the nonbanking sector is satisfied to hold exactly the amount of money that exists, no more and no less: $M^d = M^s$. The system represented by these three equations is shown in Figure 14. The determination of income in the system is shown by assuming $M^s = $25,000,000 and $k = 1/4$. The amount of money demanded is equal to supply when income is $100,000,000. A reduction of the money supply to $20,000,000 will cause income to decline to a level of $80,000,000 per year.

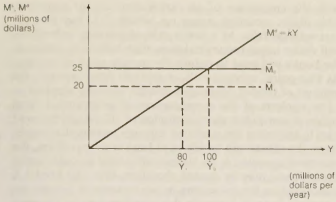


Figure 14: Relation between money demand and income (see text).

Changes in cash balances

Figure 14 shows what will happen if income temporarily exceeds the figure of $100,000,000 per year. To the right of Y_0, the amount of money demanded exceeds the existing stock of it. The way for an individual to build up his cash balance is to reduce his disbursements below his receipts. But his spending (to the extent that it is spending on final goods at least) is somebody else's income. A general attempt to build up cash balances cannot succeed—it does not induce an increase in the money supply in this model—because it will result in a decline of income throughout the system. This decline will continue to whatever level is required to make the nonbanking sector bring the amount of money it demands into line with the amount in existence. *An excess demand for money is associated with falling income.* Similarly, if the amount of money demanded falls short of the amount supplied, an individual may decide to reduce his cash balance by increasing his disbursements—but the money stays in the system; incomes will rise all around. *An excess supply of money is associated with rising income.*

The stabilization policy that this model suggests is obvious: if the relationship between income and the demand for money is stable, the system can be maintained in equilibrium by keeping the money supply constant or, in a growing economy, by allowing the money stock to grow at roughly the same rate as real output. If the relationship between income and the demand for money is found to shift about over time, the money stock should be made to grow more rapidly in periods of increasing demand for money and more slowly in periods of decreasing demand.

Comparisons of the income and money models. Although the two models seem to have nothing in common—the crucial variables of one do not even appear in the other—their descriptions of what happens during income level movements are not contradictory. Falling income is associated with an excess supply of goods and services in the income model, with an excess demand for money in the money model. Rising income is associated with an excess demand for goods in the first model, with an excess supply of money in the other. Evidently the two models give only *partial* descriptions of what is going on: one model looks at the process from the "real" side only and the other from the "monetary" side. But an excess demand for goods on one side will be associated with an excess supply of moneyon the other, and vice versa, so in this respect the two are consistent.

The controversy between the two schools of thought represented by the models has mainly to do with two issues. One issue is which set of policy instruments—fiscal or monetary—provides the best means of stabilizing the economy. The other, more fundamental, issue concerns the causes of income movements. As seen above, changes in investment were the main cause of income movements in the income model; changes in the money stock were the main cause in the money model. Simplistic as the two models are, they embody the conflicting hypotheses of the two contending schools. Income–expenditure theorists attribute the instability of income primarily to events that influence the business sector's expectations with regard to the profitability of new investment, thus influencing investment. The modern quantity theorists see the irregular time path of the money stock as the most important factor.

The gross features of economic history do not contradict either hypothesis. Private investment has indeed been the most volatile component of Gross National Product. Similarly, the movements of the money stock have conformed to those of money income: rapid inflation has been associated with a rapid growth of the money supply; severe recessions, with a decline in the money supply; and mild recessions, with a slowdown in the growth of the money supply. ("Mild" recessions may be thought of as recessions during which total employment stagnates, and the growth in unemployment, therefore, is largely due to the growth of the labour force.) The controversy has in large measure come to concern the *direction of causation:* one side maintains that shifts in investment cause income changes and infers that these in turn induce changes in the money stock which go in the same direction; the other side maintains that changes in the size or rate of growth of the money stock cause income changes that in turn will tend to fall most heavily on the investment component of income.

The question of causation

The problem of resolving this controversy is twofold. First, the theoretical issue is less clear-cut than implied above. Each side acknowledges that neither investment nor the money supply is autonomous and that each affects the other. The question has become, therefore, which model is "most nearly true" and which model, consequently, should be regarded as a "first approximation" in guiding stabilization policy.

Second, the empirical methods at the disposal of economists are not yet adequate for settling such issues. Attempts have been made to compare the performance of the two models by testing whether the best predictions of income are obtained by using actual data for "autonomous expenditures" and assuming that consumption will obey the consumption–income relation that has generally obtained in the past or by using actual money stock figures and assuming that money demand will obey the relation to income that has generally obtained in the past. These attempts have bogged down in disagreements on various statistical matters and must be judged inconclusive. They have shown, however, that even with consumption functions and money demand functions that are a good deal more "reasonable" than the naive relationships above, the predictions of both models are too inaccurate for the purposes of stabilization policy.

Each model emphasizes one set of disturbances ("real" or "monetary", respectively) that will cause income to change. Each gives a partial view of the process of income-level movements. What is needed, therefore, is a third model explaining the linkages between "real" and "monetary" forces that these two simple models leave out.

Interest-rate policy. The third model brings a crucially important—but hitherto generally neglected—element into the picture of the economic system; namely, financial markets. For simplicity, the model has only one financial market; there is only one class of financial instruments (referred to as "securities") and only one yield (a single interest rate). The standard security may be thought of as a bond promising to pay annually a fixed number of dollars. The interest rate is the value of the coupon expressed as a percentage of the market price of the bond. Consequently, if excess demand for bonds brings their price up, the interest rate falls; if excess supply sends the bond price down, the interest rate rises.

Role of financial markets

The working of the financial market is depicted in the model as follows. Investment by the business sector is assumed to be financed through the issue of securities. The higher the interest rate that firms must pay on their securities, the smaller will be the investment program that they see as promising to be profitable. Thus investment will be discouraged by a rise and encouraged by a fall in the interest rate. Households, in deciding how to divide their income between consumption and saving, will consider the amount of future consumption that can be gained by abstaining from consumption now (*i.e.*, by saving). The higher the rate of interest, the larger the amount that can be spent on future consumption per dollar not spent in the present. Thus saving is encouraged by a rise and discouraged by a fall in the interest rate. Coins, notes, and some checking deposits are assets on which interest is not

paid. An individual who holds them has the alternative of converting some part of his money holdings into interest-bearing form. Thus the amount of money demanded will tend to diminish when the interest rate rises and to increase when it falls. The banking system creates money by buying assets from the public, paying for the assets through the issuance of additional monetary liabilities (*e.g.*, checking deposits). Banks must decide whether turning part of their cash reserves to an income-earning use is worth the risks of decreased "liquidity" entailed by lower bank reserves. Hence there is a tendency for the money supply to increase when the interest rate rises and to decrease when it falls.

Interest rate as price

In this model, then, the interest rate acts as a price in controlling the behaviour of the individual agents whose activities are to be coordinated. The interest rate itself is determined by the demand for and supply of money and securities. An increase in planned investment will be associated with the issuance of a large volume of securities. It will tend, therefore, to create an excess supply of securities, to lower securities prices, and to raise the rate of interest. Similarly, an increase in planned saving will tend to create an excess demand for securities, to raise their prices, and to lower the rate of interest. An increased demand for money will, in part, reduce the demand for and increase the supply of securities; it tends to create an excess supply of securities and to raise the interest rate. An increase in the supply of money will tend to reduce the rate of interest.

These qualitative propositions are the framework of the new model, integrating the two previous models as follows: (1) $I = I(r)$; (2) $C = C(Y,r)$; (3) $S = Y - C$; (4) $S = I$; (5) $M^d = M^d(Y,r)$; (6) $M^s = M^s(r)$; and (7) $M^d = M^s$. Here, Equations 1 through 4 restate the income model with the modification that investment is no longer simply "autonomous" but depends on the current level of the interest rate (r). Equations 5 through 7 restate the money model with the modification that the demand for money and the supply of money also depend on the interest rate. Two conditions now have to be simultaneously fulfilled for the system to be in equilibrium: desired saving must equal desired investment (Equation 4), and the amount of money that individuals and firms desire to hold must equal the amount that the banking sector desires to supply (Equation 7).

Only a partial account of the ways in which this model works can be given here. The following illustrative examples begin with the system in equilibrium at full employment. The first illustration adopts the view of someone who has learned the income model and hence is thoroughly imbued with the idea that rising income results from an excess of planned investment over planned saving. Faced with the proposition, drawn from the money model, that an increase in the money supply will also cause income to rise, he will ask how such a change in the money supply can cause a discrepancy between saving and investment when there was none to begin with. The answer is that an increase in M^s will mean that there is an excess supply of money and a corresponding excess demand for commodities and securities, but the *immediate* impact of excess demand will be felt almost exclusively in the securities market. The excess demand for securities drives the rate of interest down—and this encourages investment and discourages saving. At that point, consequently, a "gap" opens up between desired saving and investment.

For the second illustration, consider instead someone who has learned the money model and who, consequently, knows that income falls when the amount of money demanded exceeds the supply. In Keynes's work the "disturbance" given the most play is some unspecified event that makes business firms take a darker view of the returns to be expected from new investment. Hence, the amount of investment that they will want to undertake at the prevailing interest rate declines. The question is how such a change in planned investment can cause a discrepancy between money demand and money supply when there was none to begin with. The simplest answer is that a decline in planned investment will be associated with a reduction in the amount of securities floated on the market and thus with the emergence of an excess demand for securities. This drives securities prices up, which is to say that the interest rate falls. At a lower rate of interest, individuals will desire larger money balances than before; in addition, the banks will tend to reduce the money stock somewhat. At that point, consequently, a gap will open between the amount of money demanded and the amount supplied.

Government fiscal policy

The analysis of the consequences of government fiscal action is somewhat more complicated. If the government tries to stimulate the economy through increased expenditures, the effects will be felt in at least two ways. First, the increased spending is an "injection" added to commodity demand and may be treated, therefore, from the Model A standpoint in the same way as an increase in private investment. Second, however, this spending may be financed through increased taxes, through government borrowing, through creation of new money, or through some combination of the three. The strongest effects are gained by following the third alternative, the creation of new money. The excess demand for goods and services created by the increase in spending will then be matched by an excess supply of money, which, as seen above, will drive down the interest rate and cause increased investment, etc. To the direct stimulus of the spending program, this method of paying for it adds the indirectly achieved stimulus of increased private investment. (Needless to say, the double effect on money income is not always desirable. The fact that this method of financing government spending has almost always been heavily resorted to in wartime accounts for the historical association of large inflations with wars.) The method of the second alternative, government borrowing, consists of financing the increase in spending through the issue of government bonds. This creates an excess supply of securities, driving up the interest rate. At the higher interest rate, money demand is lessened and money supply somewhat increased, but the consequent excess supply of money will be of smaller magnitude than that entailed by creating new money. The higher interest rate will also discourage private investment. Thus the indirect effects of government borrowing are seen to involve a decrease in private investment partially offsetting the initial increase in government spending. The size of this offset has become one of the major issues between "monetarist" and "income–expenditure" economists. The monetarists argue that the offset is so nearly complete that fiscal action will be largely ineffectual unless it is accompanied by an increase in the money supply, but an increase in the money supply will have almost as powerful effects without any simultaneous fiscal action. The other side concedes that fiscal action will be more powerful when financed through changes in the money supply but maintains that countercyclical variations in government spending financed through borrowing must still be regarded as an important stabilization method.

The "natural" rate of interest and effective demand. *The thought of Knut Wicksell.* Around the turn of the century, the Swedish economist Knut Wicksell contributed greatly to the understanding of the function of the rate of interest in the mechanism determining income and price-level movements. Assuming an economy initially in full-employment equilibrium, Wicksell analyzed the various ways in which the system might depart from that position because of discrepancies between the prevailing market rate of interest and what he termed the "natural rate." The latter rate, hypothetical rather than directly observable, may be thought of as the interest rate level that would have to prevail for the system to remain at full employment with stable prices. In illustrating the use made of this concept, one should distinguish between processes initiated by "real" disturbances (the first two examples below) and those initiated by "monetary" disturbances (the third example).

The first example is one in which business firms see increased opportunities for profitable investment. The system is already at full employment, and hence an increase in spending on investment without a corresponding decrease in spending for consumption would spell inflation. What kind of adjustment will maintain stable prices? A rise in the interest rate will (1) moderate the increase in

investment spending and (2) cause households to divert some of their income from consumption into increased saving. The hypothetical level of the interest rate that will exactly match the net increase in investment with the decrease in consumption (increase in saving) is the new value of Wicksell's "natural rate." But the adjustment of the market rate may, for several reasons, come to a halt after going only part of the way to the new natural rate level. At some level of the market rate below natural rate, where planned investment still exceeds the savings that households provide for its financing, the banks may step in and finance the difference through expansion of the money supply. Thus inflation results. In Wicksell's theory there is inflationary pressure on the system associated with a market rate below the natural level and, in the version of it given here, with an increase in the money supply.

Inflationary response to the interest rate

The second example involves a change in public behaviour in that households desire to save more and consume less, out of any given level of income. The decreased demand for consumption goods threatens to cause deflation (or unemployment). To prevent this it is necessary to switch resources over to investment goods production, which requires a lowering of the interest rate. Thus an increase in saving means that the natural rate of interest declines. The adjustment of the market rate of interest may again be incomplete if falling rates induce banks, say, to reduce their new lending below scheduled loan repayments, thus reducing the money supply. Part of the saving done by households then goes, directly or indirectly, into reducing the private sector's indebtedness to banks rather than into financing investment. Thus deflationary pressure on the system is, in Wicksell's theory, associated with a market rate of interest above the natural rate and, in this example, with a decreased supply of money.

The third example is one in which banks desire to expand their loans and, thereby, their monetary liabilities—creating a "monetary" disturbance. Since "real" incentives to save and to invest have not changed, the natural rate of interest has not changed. The increased supply of bank credit will, however, drive the market rate down. It goes below the natural rate, the money supply is increased in the process, and inflation is the result.

Keynes and Wicksell. Keynes first took up Wicksell's idea in his *Treatise on Money* (1930). In Wicksell's writings, discrepancies between the natural and market rates had invariably been associated with expansion or contraction of bank credit. Keynes emphasized that such discrepancies may develop and continue without expansion or contraction of the money supply, because of speculation in the securities markets. For example, if the natural rate has decreased and the market rate starts to edge down in response to an excess of the household savings offered in demand for securities over the supply of new securities marketed to finance investment, securities prices will rise. This, Keynes suggested, will cause some speculators in "old" securities to enter the market and supply savers with securities from their holdings. The excess demand pressure on the market is thus relieved and the rise in prices (fall of the market rate) halted. The motive for these transactions is the speculators' hope that they can buy back their securities at lower prices later. In the meantime, the speculators hold their funds in the form of ready money; there has been an increase in the amount of money demanded rather than, as Wicksell assumed, a decrease in the money supply.

The Wicksell–Keynes theory was an important contribution to the theory of the income-determination process. Yet there is nothing in its main elements that should have startled a pre-Wicksellian traditional economist. The natural rate is essentially the interest rate that would prevail in general equilibrium, and a market rate different from the natural rate is a disequilibrium interest rate. Traditional economics was clear enough as to the consequences that will follow if one or more of the prices in the system "gets stuck" at a disequilibrium level. The Wicksell–Keynes theory, therefore, may be regarded as a particular application of previously familiar principles.

Keynes returned to the Wicksellian theme in *The General Theory of Employment, Interest and Money* (1936), but in that revolutionary work he gave the theory a genuinely novel twist: he argued that the system might be seriously out of equilibrium even though the prevailing interest rate was exactly at the Wicksellian natural level. This might happen because the interest rate mechanism cannot ensure that the plans of households and business firms with regard to future consumption and production will mesh with each other. There might, for example, be an increase in household saving—that is, a decrease in the demand for current consumption goods and an increase in the planned demand for future goods. Coordination of household and business activities requires that business firms respond by shifting resources out of the production of present consumption goods and into investment activities that lay the groundwork for increased output in the future. Households, in carrying out their saving decisions, do not place contractual orders with producers for future deliveries of particular goods and services. Thus the future demands implicit in current saving decisions may not be effectively communicated to producers, as efficient coordination would require. If producers draw up their investment plans on the basis of forecasts of future demand that do not correspond to the spending that households are prepared to undertake in the future, there will be an excess demand (or excess supply) for future output.

Coordination of consumer and producer planning

Such effective demand failure is not the result of changes in interest rates or in the supply of money. The logical way of dealing with it—when it occurs—is through fiscal policy measures. The effective demand doctrine is the signal contribution of Keynesian economics to income and employment theory. It is thus no coincidence that Keynesian economics has become associated with an emphasis on the use of fiscal, rather than monetary, stabilization policies.

(Ed.)

BIBLIOGRAPHY

General works. Comprehensive resources include JOHN EATWELL, MURRAY MILGATE, and PETER NEWMAN (eds.), *The New Palgrave: A Dictionary of Economics*, 4 vol. (1987); and DOUGLAS GREENWALD (ed.), *The McGraw-Hill Encyclopedia of Economics*, 2nd ed. (1994). (Ed.)

Utility and value. Two good introductory discussions are ROBERT DORFMAN, *The Price System* (1964), and *Prices and Markets*, 3rd ed. (1978). JOSEPH SCHUMPETER, *History of Economic Analysis*, ed. by ELIZABETH BOODY SCHUMPETER (1954, reissued 1986), is a classic work. An excellent brief discussion can be found in GEORGE J. STIGLER, *Essays in the History of Economics* (1965, reprinted 1987), especially essays 5, 6, and 12. Three rather advanced works on modern value theory are J.R. HICKS, *Value and Capital*, 2nd ed. (1950, reissued 1974); PAUL A. SAMUELSON, *Foundations of Economic Analysis*, enlarged ed. (1983); and J. HÜSLER and R.-D. REISS (eds.), *Extreme Value Theory* (1989), containing conference proceedings. MARC R. TOOL, *Essays in Social Value Theory: A Neo-Institutionalist Contribution* (1986), provides very insightful views.

Seminal works in the history of value theory include DAVID RICARDO, *On the Principles of Political Economy and Taxation* (1817, reissued 1981); F.Y. EDGEWORTH, *Mathematical Psychics* (1881, reprinted 1967); CARL MENGER, *Principles of Economics* (1950, reissued 1981; originally published in German, 1871); LÉON WALRAS, *Elements of Pure Economics; or, The Theory of Social Wealth* (1954, reprinted 1984; originally published in French, 1874); W. STANLEY JEVONS, *The Theory of Political Economy*, 5th ed. (1957); ALFRED MARSHALL, *Principles of Economics*, 9th ed., 2 vol. (1961), also discussing price; JOHN WEEKS, *Capital and Exploitation* (1981), a study of Marx's labour theory of value; HERMANN HEINRICH GOSSEN, *The Laws of Human Relations and the Rules of Human Action Derived Therefrom* (1983; originally published in German, 1854); and K.K. VALTUKH, *Marx's Theory of Commodity and Surplus-Value: Formalised Exposition*, trans. from Russian (1987). Two good discussions of the more recently popular theories of utility are DAVID M. KREPS, *Notes on the Theory of Choice* (1988); and BILL GERRARD (ed.), *The Economics of Rationality* (1993).

Among the best modern textbooks in microeconomics are DAVID D. FRIEDMAN, *Price Theory*, 2nd ed. (1990); DAVID M. KREPS, *A Course in Microeconomic Theory* (1990); and HAL R. VARIAN, *Microeconomic Analysis*, 3rd ed. (1992), and *Intermediate Microeconomics*, 3rd ed. (1993). (W.J.B./Ed.)

Price. Major historical treatises include ADAM SMITH, *An Inquiry into the Nature and Causes of the Wealth of Nations*, 2 vol. (1776, reissued in 1 vol., 1991); and JOHN STUART MILL, *Principles of Political Economy*, 2 vol. (1848, reissued in 1 vol., 1994). Among the best contemporary textbooks are GEORGE J.

STIGLER, *The Theory of Price*, 4th ed. (1987); F.M. SCHERER and DAVID ROSS, *Industrial Market Structure and Economic Performance*, 3rd ed. (1990); PAUL A. SAMUELSON and WILLIAM D. NORDHAUS, *Economics*, 14th ed. (1992); and WILLIAM J. BAUMOL and ALAN S. BLINDER, *Economics*, 6th ed. (1994), and *Macroeconomics*, 6th ed. (1994). FRANK H. KNIGHT, *The Economic Organization* (1933, reissued 1967), presents a classic statement of the functions of an economic system. Selected applied analyses include MILTON FRIEDMAN and ROSE D. FRIEDMAN, *Capitalism and Freedom* (1962, reissued 1982), especially chapters 6–12; JOE S. BAIN, *Essays on Price Theory and Industrial Organization* (1972); REUBEN A. KESSEL, R.H. COARSE, and MERTON H. MILLER (eds.), *Essays in Applied Price Theory* (1980); ARTHUR M. OKUN, *Prices and Quantities: A Macroeconomic Analysis* (1981); and JACK HIRSHLEIFER and AMIHAI GLAZER, *Price Theory and Applications*, 5th ed. (1992).
(G.J.S./Ed.)

Market structure. Noteworthy texts include WILLIAM FELLNER, *Competition Among the Few: Oligopoly and Similar Market Structures* (1949, reissued 1965); JOE S. BAIN, *Barriers to New Competition: Their Character and Consequences in Manufacturing Industries* (1956, reprinted 1993), and *Industrial Organization*, 2nd ed. (1968), a general textbook; CARL KAYSEN and DONALD F. TURNER, *Antitrust Policy: An Economic and Legal Analysis* (1959); EDWARD CHAMBERLIN, *Theory of Monopolistic Competition*, 8th ed. (1962); JOAN ROBINSON, *Economics of Imperfect Competition*, 2nd ed. (1969, reissued 1976); and WILLIAM G. SHEPHERD and CLAIR WILCOX, *Public Policies Toward Business*, 6th ed. (1979), a good general textbook. JOSEPH A. SCHUMPETER, *Capitalism, Socialism, and Democracy*, 6th ed. (1987), provides a still brilliant analysis of various market structures; and MORTON I. KAMIEN and NANCY L. SCHWARTZ, *Market Structure and Innovation* (1982), follows nicely from Schumpeter's book. CHARLES E. LINDBLOM, *Politics and Markets* (1977), is also a classic and an excellent examination of market structures; and RICHARD E. QUANDT and DUSAN TRISKA (eds.), *Optimal Decisions in Markets and Planned Economies* (1990), offers a more current look at themes treated by Lindblom. Special topics related to market structure can be found in DON E. WALDMAN, *Antitrust Action and Market Structure* (1978); and ELHANAN HELPMAN and PAUL R. KRUGMAN, *Trade Policy and Market Structure* (1989). Excellent treatments of industrial organization include JEAN TIROLE, *The Theory of Industrial Organization* (1988), a graduate-level text; RICHARD SCHMALENSEE and ROBERT D. WILLIG (eds.), *Handbook of Industrial Organization*, 2 vol. (1989); and DENNIS W. CARLTON and JEFFREY M. PERLOFF, *Modern Industrial Organization*, 2nd ed. (1994), an accessible intermediate-level text.
(J.S.B./Ed.)

Production. ROLF FÄRE, *Fundamentals of Production Theory* (1988), contains an excellent introductory treatment. Authoritative intermediate-level discussions may be found in WILLIAM J. BAUMOL, *Economic Theory and Operations Analysis*, 4th ed. (1977), VERNON L. SMITH, *Investment and Production* (1961), especially emphasizes the relationship between long-run costs and investment. Probably the best technical presentation is PAUL A. SAMUELSON, *Foundations of Economic Analysis*, enlarged ed. (1983). GEORGE J. STIGLER, *Production and Distribution Theories* (1941, reissued 1994), provides a discussion of the evolution of the theory; while GERHARD ROSEGGER, *The Economics of Production and Innovation: An Industrial Perspective*, 2nd ed. (1986); and FINN R. FØRSUND (ed.), *Topics in Production Theory* (1984), discuss the relationship between theory and empirical data. An interesting survey of recent work in this area is FERNANDO J. CARDIM DE CARVALHO, *Mr. Keynes and the Post Keynesians: Principles of Macroeconomics for a Monetary Production Economy* (1992).
(R.D./Ed.)

Distribution. Analyses of economic distribution appear in DAVID RICARDO, *Principles of Political Economy and Taxation* (1817, reissued 1981), the classical subsistence theory of wages; KARL MARX, *Capital*, vol. 1 (1886; originally published in German, 1867), also available in many later editions, treating the process of distribution as pure conflict; JOHN BATES CLARK, *Distribution of Wealth* (1899, reissued 1965), the classic work on marginal productivity theory; FRANK H. KNIGHT, *Risk, Uncertainty, and Profit* (1921, reprinted 1985); JOSEPH SCHUMPETER, *The Theory of Economic Development* (1934, reprinted 1987; originally published in German, 1912); PAUL H. DOUGLAS, *The Theory of Wages* (1934, reissued 1964), which sets forth the famous Cobb–Douglas function; K.J. ARROW *et al.*, "Capital–Labor Substitution and Economic Efficiency," *The Review of Economics and Statistics*, 43:225–250 (1961); J.R. HICKS, *The Theory of Wages*, 2nd ed. (1963, reissued 1973), a sophisticated treatment of marginal productivity theory; and NICHOLAS KALDOR, "Alternative Theories of Distribution," in his *Essays on Value and Distribution*, 2nd ed. (1980), a discussion of various theories from Ricardo to Keynes. DAN USHER, *The Economic Prerequisite to Democracy* (1981), suggests that democracy requires broad agreement on how an economic system will distribute wealth. Other works in this area are ALAN S. BLINDER, *Toward an Economic Theory of Income Distribution* (1974); and RONALD G. EHRENBERG and ROBERT S. SMITH, *Modern Labor Economics: Theory and Public Policy*, 5th ed. (1994).
(K.E.Bo./P.L.Kl./H.O.Sc./J.P./Ed.)

Consumption. The two classic references for understanding the modern theory of consumption are FRANCO MODIGLIANI and RICHARD BRUMBERG, "Utility Analysis and the Consumption Function: An Interpretation of Cross-Section Data," in KENNETH K. KURIHARA (ed.), *Post-Keynesian Economics* (1954, reissued 1993), pp. 388–436; and MILTON FRIEDMAN, *A Theory of the Consumption Function* (1957). JOHN KENNETH GALBRAITH, *The Affluent Society*, 4th ed. (1984), studies the social and economic consequences of advanced industrialism. THORSTEIN VEBLEN, *The Theory of the Leisure Class* (1899, reissued 1994), argues that the consumption of the wealthy classes is aimed mainly at demonstrating status rather than satisfying basic needs. E.J. MISHAN, *The Costs of Economic Growth*, rev. ed. (1993), critiques the goal of economic growth as a major aim of policy. VANCE PACKARD, *The Hidden Persuaders*, rev. ed. (1980), provides a somewhat partisan attempt to show that advertising techniques create artificial wants through psychological manipulation. A textbook covering modern consumption theory at an advanced level is ANGUS DEATON and JOHN MUELLBAUER, *Economics and Consumer Behavior* (1980); and an excellent survey on recent contributions to this area is ANGUS DEATON, *Understanding Consumption* (1992).
(J.A.C.B./Ed.)

Economic fluctuations. Good introductions to the study of business cycles include ERIK LUNDBERG (ed.), *The Business Cycle in the Post-War World* (1955, reprinted 1986); and ALVIN HARVEY HANSEN, *Business Cycles and National Income*, expanded ed. (1964). Famous surveys of business cycle theories are JOSEPH A. SCHUMPETER, *Business Cycles*, 2 vol. (1939, reprinted 1982); and GOTTFRIED HABERLER, *Prosperity and Depression*, 5th ed. (1964). J. TINBERGEN, *Statistical Testing of Business-Cycle Theories*, 2 vol. (1939, reissued 2 vol. in 1, 1968), attempts to verify by econometric analysis the theories surveyed in Haberler's work. The nontheoretical approach to business-cycle research is set forth in ARTHUR F. BURNS and WESLEY C. MITCHELL, *Measuring Business Cycles* (1946); and further developed in GEOFFREY HOYT MOORE, *Business Cycle Indicators*, 2 vol. (1961). History and politics are dealt with in VIVIAN WALSH and HARVEY GRAM, *Classical and Neoclassical Theories of General Equilibrium* (1980); DENNIS C. MUELLER, *Public Choice II* (1989); and JAMES E. ALT and KENNETH A. SHEPSLE (eds.), *Perspectives on Positive Political Economy* (1990).

Two good introductions to macroeconomics are GEORGE T. MCCANDLESS, JR., *Macroeconomic Theory* (1991), neoclassical in approach; and JOSEPH STIGLITZ, *Economics* (1993), Keynesian-oriented. At the intermediate level, the two competing alternatives for reaching a sound understanding of national income theory are ROBERT J. BARRO and VITTORIO GRILLI, *European Macroeconomics* (1994), which applies the intertemporal equilibrium approach to macroeconomic analysis; and RUDIGER DORNBUSCH and STANLEY FISCHER, *Macroeconomics*, 6th ed. (1994), which follows an IS-LM approach. Three advanced textbooks in macroeconomic theory are THOMAS J. SARGENT, *Macroeconomic Theory*, 2nd ed. (1987), and *Dynamic Macroeconomic Theory* (1987); and OLIVIER JEAN BLANCHARD and STANLEY FISCHER, *Lectures on Macroeconomics* (1989).

More specialized or intensive treatments of macroeconomics are JOHN MAYNARD KEYNES, *The General Theory of Employment, Interest, and Money* (1936, reissued 1991), the classic theoretical work in the field; SEYMOUR E. HARRIS (ed.), *The New Economics: Keynes' Influence on Theory and Public Policy* (1947, reprinted 1973); and GARDNER ACKLEY, *Macroeconomic Analysis and Theory* (1978), an introductory text. Later evaluations by leading economists of the significance and influence of Keynesian ideas may be found in ROY F. HARROD, *The Life of John Maynard Keynes* (1951, reissued 1982), offering insight into the genesis of Keynes's ideas; ROBERT LEKACHMAN (ed.), *Keynes' General Theory: Reports of Three Decades* (1964); AXEL LEIJONHUFVUD, *On Keynesian Economics and the Economics of Keynes* (1968); and HERBERT STEIN, *The Fiscal Revolution in America*, rev. ed. (1990).

Some important perspectives on the successes and failures of economic strategies that have resulted in fluctuations are JOHN KENNETH GALBRAITH, *Economics and the Public Purpose* (1973), and *Economics in Perspective: A Critical History* (1987); and DOUGLASS C. NORTH, *Institutions, Institutional Change, and Economic Performance* (1990). Two other works that offer perspectives on economic theory and its applications are NEIL DE MARCHI and MARK BLAUG (eds.), *Appraising Economic Theories* (1991); and MARK BLAUG, *The Methodology of Economics, or, How Economists Explain*, 2nd ed. (1992).
(H.Gu./Ed.)

Ecuador

Ecuador (in full, Republic of Ecuador; Spanish: República del Ecuador), a country of northwestern South America, straddles part of the Andes Mountains and occupies part of the Amazon basin. Lying on the Equator, from which its name derives, it borders Colombia to the north, Peru to the east and the south, and the Pacific Ocean to the west; it includes the Pacific island group of the Galápagos Islands, or the Archipiélago de Colón. It has an area of 103,930 square miles (269,178 square kilometres)—more than twice the size of Nicaragua—including 100,844 square miles on the South American continent. A series of border disputes with Peru, which erupted periodically into warfare during the 20th century, were resolved by treaty in 1998. The capital, Quito, is located in the Andean highlands in the north-central part of the country.

Ecuador is one of the most environmentally diverse countries in the world, and it has contributed notably to the environmental sciences. The first scientific expedition to measure the circumference of the Earth, led by Charles-Marie de La Condamine, was based in Ecuador; research in Ecuador by renowned naturalists Alexander von Humboldt and Charles Darwin helped establish basic theories of modern geography, ecology, and evolutionary biology. Ecuador has a deeply ingrained cultural heritage. Much of what is now Ecuador came to be included in the Inca empire, the largest political unit of pre-Columbian America. Economically, Ecuador has become known for exporting (erroneously named) Panama hats and agricultural products, notably bananas. Its history has been marked by political and economic challenges, including long periods of military rule, boom-and-bust economic cycles, and inequitable distributions of wealth. Ecuador is unusual among Latin American nations in having two major centres of population and commerce—with the vibrant port city of Quayaquil acting as a counterbalance to Quito.

This article focuses on the land and people of continental Ecuador; for information on the Galápagos Islands, see the *Micropædia:* GALAPAGOS ISLANDS.

This article is divided into the following sections:

Physical and human geography 953
- The land 953
  - Relief
  - Drainage
  - Soils
  - Climate
  - Plant and animal life
  - Settlement patterns
- The people 956
  - Ethnic and linguistic composition
  - Religion
  - Demographic trends
- The economy 957
  - Resources
  - Agriculture
  - Mining and industry
  - Finance and trade
  - Transportation
- Administration and social conditions 958
  - Government
  - Education
  - Health and welfare
- Cultural life 959

History 960
- Pre-Spanish era 960
- The colonial period 960
- Early national history, 1830–*c.* 1925 960
  - Liberal-Conservative hostilities
  - Rivalry between Flores and Rocafuerte (1830–45)
  - Breakdown of national government (1845–60)
  - The regime of García Moreno (1860–75)
  - Shift to liberalism (1875–97)
  - Problems of the early 20th century
- Modern history 961

Bibliography 962

Physical and human geography

THE LAND

Relief. The Ecuadoran mainland is divided into three main physical regions: the Costa (coastal region), the Sierra (highland region), and the Oriente (eastern region, also called the Amazon region). The Costa is composed of lowlands that extend eastward from the Pacific Ocean to the western edge of the Andes and rise from sea level to an elevation of 1,650 feet (500 metres). Running north-south, small coastal mountain ranges—the Colonche, Chindul, and Mache mountains—rise to 2,600 feet. Between these coastal ranges and the Andes, interior valleys are mantled with silt deposits left by rivers that largely drain into the Gulf of Guayaquil. Puná, in the gulf, is the major island.

The Costa

The western and central ranges of the Andes bordering the Sierra constitute the country's highest and most continuous mountain chains. Many peaks are volcanic or snow-covered; these include Cayambe, 18,996 feet (5,790 metres); Antisana, 18,714 feet (5,704 metres); Cotopaxi—the world's highest active volcano—19,347 feet (5,897 metres); Chimborazo, 20,702 feet (6,310 metres); Altar, 17,451 feet (5,319 metres); and Sangay, 17,158 feet (5,230 metres). The two ranges are connected at intervals by transversal mountain chains, between which are large, isolated valleys or basins, called *hoyas.*

The Oriente begins with the eastern spur of the central range, which extends to the border with Peru. This region is crossed by the eastern—and least important—cordillera of the Andes, also composed of three sections: the Cordillera de Galeras, which includes the northern mountains and such peaks as Reventador (11,434 feet) and Sumaco (12,759 feet); the Cordillera de Cutucú, which borders the Upano valley and includes the central peaks; and the Cordillera del Cóndor to the south, which borders the Zamora valley. Beyond this eastern cordillera, to the east, is the Amazon basin, extending below 900 feet.

The volcanic Galápagos Islands consist of 19 rugged islands and scores of islets and rocks situated about 600 miles west of the mainland. The largest island, Isabela (Albemarle), rises to 5,541 feet (1,689 metres) at Mount Azul, the archipelago's highest point. The second largest island is Santa Cruz.

Drainage. Numerous rivers originate in the mountains, pass through the hoyas of the Sierra, and flow either west to the Pacific coast or east to the Amazon River. The main watercourse of the Costa is the Guayas River. Formed by the juncture of the Daule and Babahoyo rivers and their affluents, the Guayas River is navigable for the greater part of its course. Other rivers that flow to the ocean include the Santiago, the Cayapas, the Esmeraldas, and the Naranjal.

In the Sierra the rivers, which are torrential in their upper courses, become calmer in the plains areas but remain, nonetheless, unnavigable.

The rivers of the Oriente carry the greatest volume of water. The most important is the Napo River, which receives the Coca and Aguarico rivers as well as other large tributaries as it takes its course toward Peru, where it joins the Amazon River. Other large rivers include the Pastaza, Morona, and Santiago, all of which drain into the Marañón River in Peru.

Soils. Ecuador's soils are among the most varied on



MAP INDEX

Political subdivisions

Azuay, *province* . . . 3 05 S 79 20 W
Bolívar, *province* . . . 1 35 S 79 05 W
Cañar, *province* . . . 2 30 S 79 00 W
Carchi, *province* . . . 0 45 N 78 05 W
Chimborazo, *province* . . . . . 1 55 S 78 45 W
Cotopaxi, *province* . . . . . 0 50 S 78 50 W
El Oro, *province* . . . 3 30 S 79 50 W
Esmeraldas, *province* . . . . . 0 50 N 79 15 W
Galápagos, *province* . . . . . 0 00 90 30 W
Guayas, *province* . . . . . 2 00 S 80 00 W
Imbabura, *province* . . . . . 0 22 N 78 25 W
Loja, *province* . . . 4 10 S 79 30 W
Los Ríos, *province* . . . . . 1 25 S 79 35 W
Manabí, *province* . . . 0 40 S 80 05 W
Morona-Santiago, *province* . . . . . 2 30 S 77 45 W
Napo, *province* . . . 0 25 S 76 55 W
Orellana, *province* . . . . . 1 00 S 77 30 W
Pastaza, *province* . . . . . 1 55 S 77 00 W
Sucumbíos, *province* . . . . . 0 20 S 77 25 W
Pichincha, *province* . . . . . 0 10 S 78 40 W
Tungurahua, *province* . . . . . 1 15 S 78 30 W
Zamora-Chinchipe, *province* . . . . . 4 15 S 78 50 W

Cities and towns

Ambato . . . . . 1 15 S 78 37 W
Arenillas . . . . . 3 33 S 80 04 W
Atuntaqui . . . . . 0 20 N 78 13 W
Azogues . . . . . 2 44 S 78 50 W
Babahoyo . . . . . 1 49 S 79 31 W
Bahía de Caráquez . . . . . 0 36 S 80 25 W
Balzar . . . . . 1 22 S 79 54 W
Baños . . . . . 1 24 S 78 25 W
Biblián . . . . . 2 42 S 78 52 W
Calceta . . . . . 0 51 S 80 10 W
Cañar . . . . . 2 33 S 78 56 W
Cariamanga . . . . . 4 20 S 79 33 W
Catacocha . . . . . 4 04 S 79 38 W
Catamayo . . . . . 3 59 S 79 21 W
Catarama . . . . . 1 35 S 79 28 W
Cayambe . . . . . 0 03 N 78 08 W
Celica . . . . . 4 07 S 79 57 W
Chone . . . . . 0 41 S 80 06 W
Chunchi . . . . . 2 17 S 78 55 W
Coca, see Puerto Francisco de Orellana
Cotacachi . . . . . 0 18 N 78 16 W
Cuenca . . . . . 2 53 S 78 59 W
Daule . . . . . 1 52 S 79 58 W
El Ángel . . . . . 0 37 N 77 56 W
El Carmen . . . . . 0 16 S 79 26 W
El Empalme, see Velasco Ibarra
El Guabo, see Guabo
Esmeraldas . . . . . 0 59 N 79 42 W
General Leonidas Plaza Gutiérrez . . . 2 58 S 78 25 W
Girón . . . . . 3 10 S 79 08 W
Guabo (El Guabo) . . . 3 15 S 79 51 W
Gualaceo . . . . . 2 54 S 78 47 W
Gualaquiza . . . . . 3 24 S 78 33 W
Guamote . . . . . 1 56 S 78 43 W
Guano . . . . . 1 35 S 78 38 W
Guaranda . . . . . 1 36 S 79 00 W
Guayaquil . . . . . 2 10 S 79 54 W
Huaquillas . . . . . 3 29 S 80 14 W
Ibarra . . . . . 0 21 N 78 07 W
Jipijapa . . . . . 1 20 S 80 35 W
Lago Agrio (Nueva Loja) . . . . . 0 06 N 76 52 W
Latacunga . . . . . 0 56 S 78 37 W
Loja . . . . . 4 00 S 79 13 W
Macará . . . . . 4 23 S 79 57 W
Macas . . . . . 2 19 S 78 07 W
Machachi . . . . . 0 30 S 78 34 W
Machala . . . . . 3 16 S 79 58 W
Manta . . . . . 0 57 S 80 44 W
Milagro . . . . . 2 07 S 79 36 W
Montecristi . . . . . 1 03 S 80 40 W
Muisne . . . . . 0 36 N 80 02 W
Naranjal . . . . . 2 40 S 79 37 W
Naranjito . . . . . 2 13 S 79 29 W
Nueva Loja, see Lago Agrio
Otavalo . . . . . 0 14 N 78 16 W
Pajan . . . . . 1 34 S 80 25 W
Pasaje . . . . . 3 20 S 79 49 W
Pelileo . . . . . 1 19 S 78 32 W
Pillaro . . . . . 1 10 S 78 32 W
Piñas . . . . . 3 40 S 79 39 W
Portoviejo . . . . . 1 03 S 80 27 W
Puerto Ayora . . . . . 0 45 S 90 19 W
Puerto Baquerizo Moreno . . . . . 0 54 S 89 36 W
Puerto Francisco de Orellana (Coca) . . . . . 0 28 S 76 58 W
Pujili . . . . . 0 57 S 78 41 W
Puyo . . . . . 1 28 S 77 59 W
Quevedo . . . . . 1 02 S 79 27 W
Quito . . . . . 0 13 S 78 30 W
Riobamba . . . . . 1 40 S 78 38 W
Rocafuerte . . . . . 0 55 S 80 26 W
Rosa Zárate . . . . . 0 20 N 79 28 W
Salinas . . . . . 2 13 S 80 58 W
Salitre . . . . . 1 50 S 79 48 W
Samborondón . . . . . 1 57 S 79 44 W
San Gabriel . . . . . 0 36 N 77 49 W
San José de Chimbo (San José) . . . . . 1 41 S 79 02 W
San Lorenzo . . . . . 1 17 N 78 50 W
San Miguel . . . . . 1 42 S 79 02 W
San Miguel de Salcedo . . . . . 1 02 S 78 34 W
Sangolquí . . . . . 0 19 S 78 27 W
Santa Ana . . . . . 1 13 S 80 23 W
Santa Elena . . . . . 2 14 S 80 51 W
Santa Isabel . . . . . 3 16 S 79 19 W
Santa Rosa . . . . . 3 27 S 79 58 W
Santo Domingo de los Colorados . . . 0 15 S 79 09 W
Saquisilí . . . . . 0 50 S 78 40 W
Sigsig . . . . . 3 03 S 78 48 W
Sucre . . . . . 1 16 S 80 26 W
Suróa . . . . . 2 28 S 78 10 W
Tena . . . . . 0 59 S 77 49 W
Tulcán . . . . . 0 48 N 77 43 W
Valdez . . . . . 1 15 N 79 00 W
Velasco Ibarra (El Empalme) . . . . 1 03 S 79 37 W
Ventanas . . . . . 1 27 S 79 28 W
Vinces . . . . . 1 33 S 79 44 W
Yaguachi Nuevo . . . . . 2 07 S 79 41 W
Yantzaza . . . . . 3 51 S 78 45 W
Zamora . . . . . 4 04 S 78 58 W
Zaruma . . . . . 3 41 S 79 37 W

Physical features and points of interest
Abingdon, see Pinta Island
Academy Bay . . . . . 0 45 S 90 17 W
Aguarico, *river* . . . . 0 59 S 75 11 W
Altar Volcano . . . . . 1 41 S 78 24 W
Ancón de Sardinas Bay . . . . . 1 30 N 79 00 W
Andes Mountains . . 2 00 S 79 00 W
Antisana Volcano . . 0 30 S 78 08 W
Azul, Mount, *volcano* . . . . . 0 54 S 91 21 W
Cayambe Volcano . . . . . 0 00 N 77 59 W
Cayambe-Coca Ecological Reserve . . . . . 0 00 77 45 W
Cayapas, *river* . . . . . 1 05 N 79 03 W
Charles, see Santa María Island
Chatham, see San Cristóbal Island
Chimborazo Volcano . . . . . 1 28 S 78 48 W
Chindul Mountains . . . . . 0 07 N 79 46 W
Coca, *river* . . . . . 0 29 S 76 58 W
Colón, see Galápagos Islands
Colonche Mountains . . . . . 2 00 S 80 20 W
Cóndor, Cordillera del . . . . . 4 00 S 78 30 W
Costa, *region* . . . . . 1 00 S 80 00 W
Cotachi-Cayapas Ecological Reserve . . . . . 0 35 N 78 35 W
Cotopaxi Volcano . . 0 40 S 78 26 W
Curaray, *river* . . . . . 1 30 S 76 30 W
Cutucú, Cordillera de . . . . . 2 45 S 78 00 W
Daule, *river* . . . . . 2 10 S 79 52 W
Daule-Peripa Reservoir . . . . . 0 50 S 79 45 W
Duncan, see Pinzón Island
Esmeraldas, *river* . . . 0 58 N 79 38 W
Española (Hood) Island . . . . . 1 23 S 89 39 W
Fernandina (Narborough) Island . . . . . 0 25 S 91 30 W
Galápagos (Colón) Islands . . . 0 00 90 30 W
Galápagos National Park . . . . . 1 00 S 90 30 W
Galeras, Cordillera de . . . . . 0 50 S 77 32 W
Genovesa Island . . . 0 20 N 89 58 W
Guayaquil, Gulf of . . . . . 3 00 S 80 30 W
Guayas, *river* . . . . . 2 36 S 79 54 W
Hood, see Española Island
Indefatigable, see Santa Cruz Island
Isabela Island . . . . . 0 30 S 91 04 W
Jambeli Canal, *marine channel* . . . 2 55 S 79 58 W
James, see San Salvador Island
Jubones, *river* . . . . . 3 13 S 79 57 W
Marchena Island . . . 0 19 N 90 29 W
Napo, *river* . . . . . 1 00 S 75 10 W
Narborough, see Fernandina Island
Oriente, *region* . . . . . 2 00 S 77 00 W
Pacific Ocean . . . . . 1 00 S 83 00 W
Pastaza, *river* . . . . . 4 55 S 76 24 W
Pinta (Abingdon) Island . . . . . 0 35 N 90 44 W
Pinzón (Duncan) Island . . . . . 0 36 S 90 40 W
Plata Island . . . . . 1 16 S 81 06 W
Puná Island . . . . . 2 50 S 80 08 W
Reventador Volcano . . . . . 0 03 S 77 40 W
San Cristóbal (Chatham) Island . . . . . 0 50 S 89 26 W
San Salvador (James) Island . . . . 0 16 S 90 42 W
Sangay National Park . . . . . 1 45 S 78 15 W
Sangay Volcano . . . 1 58 S 78 22 W
Santa Cruz (Indefatigable) Island . . . . . 0 38 S 90 23 W
Santa Elena, Point . . . . . 2 11 S 81 00 W
Santa Elena Peninsula . . . . . 2 15 S 80 50 W
Santa Fe Island . . . . 0 49 S 90 04 W
Santa María (Charles) Island . . . 1 17 S 90 26 W
Sumaco Volcano, *mountain* . . . . . 0 34 S 77 38 W
Toachi, *river* . . . . . 0 06 N 79 13 W
Upano, *river* . . . . . 2 43 S 78 15 W
Wolf Volcano . . . . . 0 02 N 91 20 W
Yasuní National Park . . . . . 1 00 S 76 00 W
Zamora, *river* . . . . . 2 59 S 78 13 W

Earth. Volcanic activity at higher elevations in the Andes has resulted in the formation of fertile volcanic and prairie soils, such as andosols and mollisols, with dark surface layers rich in organic matter. The soils are typically, however, underlain by a yellow-coloured hardpan, locally called *cangahua*, which is often exposed on eroded steeper slopes. The eroded topsoil accumulates on lower slopes and especially on flats, which form the most desirable locations for agriculture. Indigenous peoples have developed effective methods for the fertilization of these soils, including the use of native manures, the mounding of fertilizing muck from drainage ditches, the creation of raised fields, and the use of irrigation canals.

Traditional fertilization methods

In the Costa the floodplains of the Guayas and other rivers have accumulated fertile silts from the highlands. These coastal soils are of great fertility but often consist of clays that are subject to shrinking and swelling and thus present problems for construction. The effectiveness of traditional methods of managing these soils has come to be recognized, and such techniques as embanked fields for runoff management (*albarradas*) and raised fields (artificially constructed earthen platforms built on shallow lakes or marshy areas) are encouraged.

In the Amazon basin, soils have not been fully studied and mapped; nevertheless, it appears that soils there are quite diverse, including areas of fertile alluvial soil, organic soils called histosols, and more weathered tropical soils called oxisols. The latter may be used for crops with appropriate technology, such as shifting cultivation or agroforestry (crops and useful trees managed together), but some agronomists suggest that they are better utilized for timber and other renewable tropical forest products.

Climate. Because Ecuador lies on the Equator, most of the country, except in the Sierra, experiences humid tropical climates. The Oriente is influenced throughout the year by an unstable maritime tropical air mass, while the Costa is subject to greater variations associated with seasonal movements of the intertropical convergence zone and the cold Peru Current. Local convectional processes dominate the weather in the higher parts of the Andes.

The Oriente experiences fairly continuous and abundant rainfall and high temperatures. The Costa generally has a wet season in the first half of the year and a relatively dry one in the second half. In some years, warm water collects off the coast, causing the weather phenomenon known as El Niño; this can result in torrential downpours that cause devastating ecological damage on the coast and occasionally even in the highlands. In the Sierra, rains reach a maximum during the equinoxes; there is a long dry season from June to September and a shorter one from December through January.

Ecuador has a small area of truly dry climate at the Santa Elena Peninsula along the southern coast, with annual rainfall decreasing from 40 inches (1,000 millimetres) near Guayaquil to only 4 inches at Salinas. In the highlands, annual rainfall decreases toward the centres of the canyons and valleys, sometimes dropping below 20 inches or even below 10 inches. Most of the country, however, is humid, receiving more than 20 inches of rain a year. The southern coast and the highlands receive 30 to 80 inches. The wettest areas, the northern coast and the Oriente, receive 120 to 240 inches of rain.

Temperature variations

Both the Costa and the Oriente regions are warm, temperatures varying only slightly among the seasons; much wider differences occur between day and night. Average daytime high temperatures range from 84° to 91° F (29° to 33° C), while nighttime lows fall to between 68° and 75° F (20° to 24° C). As elevation increases, temperatures drop fairly predictably at a rate of about 9° to 11° F (5° to 6° C) for every 3,300 feet. Pleasantly temperate climates occur between elevations of 2,600 and 6,600 feet. At higher elevations, frost is a possibility, especially in areas of flat relief and during cloud-free nights of the dry seasons. Above elevations of 11,800 to 12,500 feet agriculture becomes increasingly difficult because of the shrinking growing season and increasing frost hazard, and above about 16,400 feet the peaks are snowcapped.

Plant and animal life. The wet lowlands of the Oriente and the northern and southeastern corners of the Costa are covered with tropical rain forest, containing various trees, lianas, and many epiphytes. This forest thickens as it approaches the zone of maximum rainfall, which occurs between 4,000 and 5,000 feet above sea level. In the Guayas River valley, the forest includes balsa, which is exploited for its light lumber; in the eastern forest the cinchona trees were a valuable source of quinine before synthetic equivalents reduced demand for it. The trees of the Costa are fast disappearing because of exploitation, while the Oriente is increasingly under attack by commercial interests seeking to establish ranches or plantations.

In the Costa between Esmeraldas and the Gulf of Guayaquil, where the climate is affected by the Peru Current, the northern rain forest gives way southward to deciduous and semideciduous woodland. There, scattered palms produce the ivory-coloured tagua nuts used in making buttons, while the hat carludovica (*Carludovica palmata*) furnishes the fibre used for Panama hats. Areas of swampy coast and the river floodplains were once covered by thick mangrove forest, but much of it has been removed to make way for shrimp aquaculture.

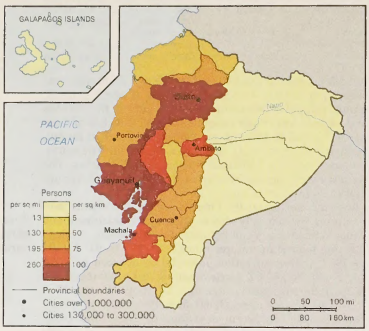
In the Sierra the decreasing growths of native vegetation of the dry valley interiors consist of a thorny woodland, which gives way toward the valley edges to a low evergreen forest and, at higher elevations, to the bunchgrasses of the high paramo. Much of the highland vegetation has been removed for agriculture or altered by periodic burning.

Rain forest fauna

In the rain forest live a wide variety of monkeys, as well as such carnivorous mammals as jaguars, ocelots, foxes, weasels, otters, skunks, raccoons, coatis (raccoon relatives), and kinkajous (tree-dwelling nocturnal animals). Hoofed mammals include the tapir, deer, and peccary. Numerous species of rodents and bats inhabit the area.

Ecuadoran bird and fish life is notably rich. Some 1,500 species of birds have been identified, different species of birds being associated with the various types of vegetation. Among many types of North American birds that migrate to Ecuador for the winter are the Virginia rail, the kingbird, the barn swallow, and the scarlet tanager. The fish population is similar to that of the Amazon River, although in the west the electric eel and the piranha are not found. All major groups of reptiles are represented.

Settlement patterns. In colonial and early modern times most people lived in the rural Sierra. By the late 20th century the growth pattern had changed, and the population majority shifted to the lowland regions, especially the Costa.



Population density of Ecuador.

In the highlands, traditional Indian and mestizo villages, hamlets, and scattered farmsteads are associated with a checkerboard pattern of small agricultural plots of corn (maize), potatoes, barley, wheat, broad beans, kidney beans, and domesticated lupine, alternating with fields temporarily lying fallow and used for grazing. Sheep are grazed on fallow land and higher-elevation pastures. Traditional housing of wattle and daub, thatch, or rammed earthen walls, with thatched roofs, has been giving way to Spanish tile or corrugated metal roofs and cement block or brick walls. Prior to the 1960s, small-scale farmers lived in a dependent relationship with large-scale haciendas, which controlled the best flat land and high pastures. Since the 1960s, land reform and economic changes have resulted in the subdivision of haciendas into more profitable medium-sized commercial farms producing dairy products, new potato varieties, fruits, and vegetables. In some cases the old rural hacienda buildings, with white walls and Spanish tile roofs, are still occupied by farm owners; in others the buildings have been abandoned by owners moving to the city or have been converted into hotels. Highland villages and towns were usually built on the Spanish colonial grid plan, which was centred on one or more plazas distinguished by church and governmental buildings.

Coastal region farming

On the coast, farmers working small plots practice a mixed tropical agriculture, growing such crops as cassava (manioc), peanuts (groundnuts), bananas, plantains, coffee, cacao, and corn; they live in houses on stilts, walled with flattened bamboo and roofed with thatch. Larger farms produce quantities of rice, cocoa, bananas, and African oil palm, while ranches raise beef cattle. Parts of the coast were colonized by mid-20th-century mestizo pioneers, especially the area to the west and northwest of Quito around Santo Domingo de los Colorados; isolated Indian populations have gradually been reduced to minority status. A similar process has been occurring in the Oriente, with oil fields and new highways allowing highland mestizos and South American Indians to move into areas settled by Amazonian Indian groups.

By the 1982 census half of the Ecuadoran population had become urban dwellers, with half of the urban population living in the two major cities. Guayaquil is the largest city, the major port and commercial centre, and also the cultural centre of the Costa. Quito, apart from its governmental activities, has become an important regional headquarters for international organizations working in the Andes and has attracted a substantial tourist trade. Other cities are much smaller, but Esmeraldas, Manta, Portoviejo, and Machala are important coastal agricultural and trade centres, and Ambato and Cuenca are the largest and most dynamic highland trade centres outside of Quito.

THE PEOPLE

Ethnic and linguistic composition. The ethnic groups of Ecuador include a number of Indian-language-speaking populations, blacks, mestizos, whites, and immigrants from a variety of foreign countries, including Lebanon, China, Korea, Japan, Italy, and Germany. Most modern censuses have not inquired about ethnicity, language, religion, or origin, so that the numbers of different groups are not precisely known. The main population components are highland and lowland mestizos.

There may have been about 700,000 Indian-language speakers as of the mid-1980s, primarily Quechua (Quichua) speakers in the Sierra. The highland Quechua speakers, many of whom are bilingual in Spanish, have only recently come to identify themselves ethnically with regions beyond their local villages; they often refer to themselves as Runa (people). They are concentrated in several distinct districts: to the north of Quito, in the vicinity of Otavalo and Cayambe; in the central highlands, from the vicinity of Latacunga to beyond the southern border of Chimborazo *provincia* and including the distinctive Salasacas Indians who live south of Ambato; in scattered areas around Cuenca in the south-central highlands; and to the north of Loja, where the Saraguro Indians live. In the southeastern lowlands is the large group of the Shuar and Achuar Indians, related to similar groups across the border in Peru; the lowland Quechua speakers (made up of several groups) occupy much of the central Amazon lowlands, along with the Huaorani (Auca), who live in a protected reserve. In the northern Oriente are the small groups of Cofán and Siona-Secoya. The Costa, from north to south, includes small Indian groups: the Awá (Kwaiker), Cayapa (Chachi), and Colorados (Tsáchila).

Black ethnic groups

The blacks of Ecuador are the descendants of slaves imported from Africa; they consist mainly of the coastal blacks of Esmeraldas *provincia* to the northeast and the highland blacks of the Chota River valley in the northern highlands. Both groups have distinctive cultures and are well-defined as ethnic groups.

Highland mestizos include numerous rural folk who occupy areas of the highlands adjacent to the Indian highland groups, as well as town and city dwellers. Mestizos of the central highlands are perhaps closest to being typical highland Ecuadorans, with little tendency to identify themselves as a distinct ethnic group or regional culture, except in the case of those living in the larger cities. Mestizos from the far southern highlands, called Lojanos, are more distinctive in lifeways and have been especially active in colonizing the Oriente and the Costa. Lowland mestizos of the Costa sharply differ from highlanders in diet, dialect, music, and identity.

Spanish is the language of business and government, although there are dialectal differences between Sierra and Costa Spanish, and Sierra Spanish has been influenced by Quechua. Most Indian males are bilingual. Several Indian languages will likely persist as mother tongues, and the concepts of bilingualism and bilingual or bicultural education are becoming increasingly important.

Religion. Ecuador is overwhelmingly Roman Catholic. The Roman Catholic church plays a significant role in

education and social services and influences the selection of significant places for festivals and pilgrimage sites, such as Quinche in the north and Biblián in the south. Protestantism continues to grow rapidly, particularly among the disadvantaged, with the largest groups being non-Pentecostal Evangelicals and Pentecostals. There is also a sizable Mormon congregation. Quito, Ambato, and Guayaquil have been urban centres of Protestant activity, and many of the Indians of the Sierra and Oriente have also converted. In Sierra *provincias* such as Carchi, Azuay, and Loja and in Manabí *provincia* in the Costa, there has been more reluctance to accept Protestant conversion. A small Jewish population is concentrated in Quito, and there are also some Bahá'í adherents.

Demographic trends. Ecuador, like other Andean countries, has experienced a population boom, the result of a decreasing death rate and a continued high birth rate. This rapid growth has resulted in a relatively young population. The country has attracted immigrants from neighbouring Colombia and from East Asia, and there has also been a small migration from Chile, including political refugees. Significant numbers of Ecuadorans have emigrated to the United States. Deprived northern highlanders tend to migrate to Quito, seeking opportunities for income not available in the countryside. The rural coastal people, on the other hand, have generally migrated to the north-central Costa and to Guayaquil, and the southern highlanders to the southeastern and northeastern Oriente and to the north-central Costa, as well as to Quito and Guayaquil. In areas where people can generate a more substantial cash income, migration has been slower.

THE ECONOMY

Ecuador is a country of enormous economic potential. Development has focused on agricultural, marine, and mineral resources, with industry playing a more limited role. The subsequent production of primary goods has been subject to cycles of boom and bust, however, and Ecuador has sought to diversify its resource exports and to seek new markets. The country has made improved standards of living, but it is still characterized by marked inequalities of wealth and well-being.

Boom and bust cycles

Resources. Ecuador's major resource is its soil, which, with its generally adequate rainfall and diverse climates, allows a wide variety of agricultural production. Particularly rich soils are found in the Guayas and other river floodplains on the coast and in the flats, floodplains, and volcanic slopes of the highlands.

The full mineral potential of Ecuador is still being discovered. There are gold deposits throughout the country and oil deposits in the northeastern Oriente. Explorations have discovered significant deposits of natural gas in the Gulf of Guayaquil, large deposits of low-grade copper ore west of Cuenca, and deposits of silver, molybdenum, iron ore, gypsum, zinc, and lead at various locations.

Forest and marine resources are also exploited. Traditional coastal dwelling construction is based on the native bamboo, and in the highlands pine and eucalyptus plantations provide fuel and construction material. A small-scale fishing industry operates mainly out of ports on the western coasts of Guayas and Manabí *provincias*. The major marine product, however, is shrimp, produced in large ponds constructed in coastal mangrove swamps. A growing problem has been the excessive destruction of these mangroves and the subsequent threat to shrimp production caused by an inadequate supply of shrimp larvae and juvenile shrimp, which are either captured in the swamps or bred by hatcheries.

The Andes Mountains present vast possibilities for hydroelectric development. The construction of larger hydroelectric plants, particularly the Agoyan and Paute projects, has notably improved hydroelectric potential, albeit with serious problems of siltation. A government agency is responsible for the development of power resources.

Agriculture. Agriculture has traditionally employed a large proportion of the population. Many Ecuadorans feed their families with the produce from their own farms; production of these subsistence crops, including corn, potatoes, beans, and cassava, is important but not accurately reflected in official figures. Commercial production of grain crops has been discouraged by imports of inexpensive grains from the United States; these imports have also encouraged a shift in diet away from traditional corn consumption and toward rice and wheat. Production of tropical specialty crops such as bananas, cacao, rice, and coffee have provided much-needed foreign exchange. Dependence on foreign imports of edible oil-producing crops and vegetable oils has been reduced through cultivation of the African oil palm. Livestock raising is widespread, with beef cattle being produced in the lowlands and dairy cattle and sheep in the highlands; chickens consume feedstuffs produced from locally grown hard corn and other crops. Pigs are raised on a small scale, but their meat does contribute to the diet, especially in the highlands. Goats are important as a source of meat in Loja *provincia* in the south, while guinea pigs are raised for food in the highlands.

Livestock raising

Mathias Oppersdorff—Photo Researchers



Farmers digging potatoes near Hacienda Zuleta in the Sierra region northeast of Quito.

Only a small proportion of the Ecuadoran territory has been reclaimed for cultivation, although unreclaimed land is valuable as forest reserves and wildlife habitats. Chemical fertilizers are mainly employed on commercial crops, while traditional farmers employ animal manures; still, overall yields could be vastly increased. Irrigation has been employed since prehistoric times in the highlands, and up to half the highland production by value is from irrigated fields; there is little further potential for expanding the highland irrigated area, however. In contrast, irrigation has been expanding rapidly on the coast.

Mining and industry. Oil and gold are the most valuable mined products. Gold has been produced in Ecuador for centuries, and much of the production comes from remote gold districts such as Nambija in Zamora *provincia*, where thousands of families live with minimal services and the miners face hazardous conditions in tunnels subject to collapse due to torrential rains. Oil extracted in the northeast and sent over the Andes via pipeline has become Ecuador's major mineral export. The state oil company operates in consortia with foreign corporations.

Industrial development is still in the early stages. Some industry is associated with the processing of primary products, including cement, refined sugar, chocolate bars, beer, pasta, bread, and instant coffee. Some import-substitute industries licensed by foreign corporations have been established, including those producing pharmaceuticals and tires and those assembling automobiles. Ecuador has had some success exporting processed foods, such as

fruit drinks and canned meats, to neighbouring countries. Ecuadoran woolen tapestries and sweaters; crafts in wood, straw, ceramics, leather, and tagua nut (called vegetable ivory); and Panama hats contribute to the economy. On a larger scale the textile and agricultural processing industries both seem promising for long-term growth.

Finance and trade. The Central Bank of Ecuador and the National Bank of Promotion, both state-controlled, have branches in all the provincial capitals. The former is the government depository and controls the monetary system, while the latter handles agricultural and industrial credit. Private or commercial banks are both foreign and domestic. The bank supervisory board is a technical organization that monitors all banking activities.

Exports include crude oil and derivatives, shrimp, bananas, coffee, and cocoa; these are sold primarily to the United States, Germany, Singapore, Panama, and Peru. Imports include machines and primary industrial materials, motor vehicles, consumer goods, and food and chemical products, bought mainly from the United States, Japan, Venezuela, Germany, Brazil, and Mexico.

Transportation. For much of its history, Ecuador relied on horse or mule transport on difficult trails or on canoe transport on coastal or Amazon river systems. Railroad development faced great difficulties, and the Quito-to-Guayaquil rail line (with a branch to Cuenca)—although locally important—is slow, antiquated, and subject to disruption by floods, landslides, and earthquakes. This is even more the case for the rail line from Quito to San Lorenzo on the coast via Ibarra. Transport was revolutionized by the paving of the Pan-American Highway, the main Ecuadoran roadway, which extends the length of the highlands from the Colombian to the Peruvian border. It is supplemented by a growing network of all-weather roads. The main highland centres are connected by asphalt roads, with asphalt or cobblestone secondary roads to regional market towns. Many rural centres are still served only by unsurfaced roads, impassable during wet periods; roads to the east of the Andes are also relatively poor. There is some concern that highway development will lead to deforestation and have adverse effects on the survival of remote Indian groups. The more likely reason for slow development, however, has been cost.

Railroad development

Goods are brought to market through labour-intensive methods by independent truckers and by peasant women and itinerant vendors, who bring small amounts of goods to market on foot, with burros or mules, or by bus. Numerous regional bus companies provide cheap, frequent, and far-ranging rural transport.

Air transport has grown, especially for the important Quito-Guayaquil connection and for international travel. The major state-controlled international airline is Compañía Ecuatoriana de Aviación, but several other international carriers serve Ecuador, landing at the major airports of Guayaquil and Quito. Other domestic airlines serve local airports, and air service to centres such as Cuenca and Machala has been established. Other air services provide access to points in the Oriente.

Guayaquil is the country's chief port, with modern facilities at Puerto Nuevo. Other modern ports include San Lorenzo, Esmeraldas, Manta, and Puerto Bolívar. Rivers, particularly in the Guayas basin, also serve as transportation arteries.

ADMINISTRATION AND SOCIAL CONDITIONS

Government. Ecuador is a sovereign democratic republic. The president and vice president are elected for four-year terms by popular, direct, and secret voting; the president may not be reelected to a second term. The chief executive is aided by a cabinet. Legislative power is vested in the unicameral National Chamber of Representatives (also called National Congress). Twelve of its members are popularly elected to four-year terms at the national level; the remaining 59 are elected to two-year terms at the provincial level.

Local government

The chief executive appoints as his representatives the provincial governors, the political chiefs in each of Ecuador's 20 *provincias,* which are divided into *cantones* (regions); these in turn are divided into *parroquias* (parishes). The chief executive also appoints the political chief in each *cantón* and a political lieutenant in each parochial district. Every province also has a provincial council presided over by the provincial prefect.

Each *cantón* constitutes a municipality, the governing of which is in the hands of a *cantón* council. The council of each provincial capital is headed by a mayor.

An array of Ecuadoran political parties draws strength from various regions, classes, ethnic groups, and professions. No party is strong throughout the country, so that alliances must be established to attain victory at the national level. Among the prominent parties is the Democratic Left party, with strength among teachers, government workers, and professionals in the more prosperous parts of the Sierra. The communist parties include the Democratic Popular Movement and the Left Broad Front, with strength in Quito and Loja, while the Ecuadoran Socialist Party has shown strength in the poorer northern and central highlands. The Popular Democracy party is a moderate party with strength in the Quito area. Centrist coastal political parties are often populist in character, associated with charismatic personalities and grass-roots political organizations. Centrist parties with strength on the coast include the Alfarista Radical Front, the Concentration of Popular Forces, and the Ecuadoran Roldosist Party. The Social Christian Party is more conservative.

Prominent political parties

Supreme Court justices are elected for terms of four years. The 16 justices are elected in theory by the National Congress, but in 1984 they were chosen by the government. There are 10 higher courts.

Education. The network of public education has been greatly expanded to promote the goal of universal literacy. Primary education is free and compulsory for six years beginning at age six. Ecuador has made progress in making education available to disadvantaged classes and ethnic groups and to women. Religious and nondenominational private schools also play a significant role. Population growth and limited funding have placed great strains on the educational system, however. Efforts are under way as well to adapt the curriculum to Ecuador's cultural diversity.

Secondary education varies from seriously overcrowded public institutions to elite private institutions emphasizing bilingualism in English, French, or German. The premier university is the Pontifical Catholic University in Quito, noted for its research programs in fields such as botany, archaeology, linguistics, and anthropology. The Polytechnic School in Quito has good programs in the sciences. There are numerous other universities with special strengths in particular areas, although the university system in general has suffered from uncertain funding and political turmoil. Many Ecuadorans seek training abroad, especially in technical fields and in business.

Universities

Much research takes place outside the universities. Geographic and environmental research and postgraduate training are conducted by the Panamerican Centre for Geographic Studies and Investigation at the Military Geographical Institute in Quito. That same building houses other environmental institutes, libraries, and laboratories. Social science institutes are also numerous, especially in Quito; they include the Andean Centre of Popular Action and a local unit of the Latin American Faculty of Social Sciences. Agricultural research is concentrated in the laboratories of the National Institute of Agricultural Investigations. Major research establishments are maintained by French and U.S. foreign assistance organizations.

Health and welfare. All public and private employees are affiliated with the National Social Security Institute. In return for a monthly deduction from employees' salaries, the agency provides such services as medical and hospital insurance coverage, state-run clinics and dispensaries, low-interest loans for surgery and mortgages, retirement pensions for civil and state employees, and pensions for widows and child dependents.

The Social Welfare Program, a division of the Ministry of Public Health, maintains public hospitals in all the provincial capitals and in the principal *cantones.* Little of the national budget is, however, devoted to public health programs, and health conditions are generally poor. A

number of endemic diseases persist, including goitre, typhoid fever, and tuberculosis.

CULTURAL LIFE

Ecuador, as discussed above, is a country of great ethnic diversity and great contrasts of wealth and poverty. People identify more with their region or village than with the country as a whole, although the government has attempted to nourish a sense of pan-Ecuadoran national identity. At a minimum the country may be divided into a dozen or so major folk-cultural regions: *norteño* mestizo, northern Quechua, central highland mestizo, Quiteño urban, central Quechua, Cuencano mestizo, Lojano mestizo, southern Quechua, Esmeraldeño black, coastal mestizo-mulatto, Shuar (Jivaro), and Amazonian Quechua. Numerous smaller or more localized cultures also exist, and there are two culturally mixed areas in the Santo Domingo and northeastern Oriente frontiers. The most prominent and representative groups are the central highland mestizos and coastal mestizo-mulatto mixed culture.

Daily life. Most Ecuadorans place great emphasis on the family, and they also create fictive kinship, which is established by the choice of godparents at baptism. Apart from baptism, important occasions in the life cycle include the 15th birthday of girls, marriage, and funerals. Many Ecuadorans make pilgrimages or dedicate themselves to the service of a particular saint. During the year, numerous religious festivals provide opportunities for parades, special food, and music and dance. Often particular holidays are associated with particular cities, such as the Day of the Dead in Ambato or Carnival in Guaranda, and they attract people from various parts of the country. The Festival of San Juan Bautista is especially important for the Indian populations of the northern highlands, who regard the holidays as an occasion for dance and music.

Festivals and holidays

Easter is an opportunity to eat *fanesca*, a soup that is close to being the Ecuadoran national dish. The soup—made of onions, peanuts, fish, rice, squash, broad beans, *chochos* (lupine), corn, lentils, beans, peas, and *melloco* (a highland tuber)—combines highland and lowland ingredients and is a culinary model of the union of diverse national characteristics. Chili sauce (*ají*) is part of most meals. Empanadas are deep-fried and stuffed savoury pastries. Typical of the coast is ceviche, made with shrimp or shellfish or even mushrooms pickled with lemon juice, cilantro (coriander), and onions. Coastal cuisine also includes deep-fried plantains and various rice dishes. Highland cuisine is based on soups and stews, including quinoa soup and barley soup, and on complex soups and stews mixing various combinations of corn, potatoes, oca, quinoa, *melloco*, beans, barley, broad beans, and squash.

Highland Indian males may wear coloured ponchos—for instance, blue in the Otavalo area and red in western Chimborazo. Traditional footwear is the sandal, and a variety of traditional hats may be worn; in some locations hair is still worn long, gathered in a ponytail. Highland Indian women may wear embroidered blouses, wrapped skirts of woolen cloth, shawls attached with a pin in front, sandals, and locally common hats or headgear.

The arts. Ecuador has a rich tradition of folk art. Quito was a colonial centre of wood carving and painting, and artisans still produce replicas of the masterpieces of the Quito school. Certain mestizo and Indian communities have specialized in particular crafts, such as agavefibre bags near Riobamba and Salcedo; wood carving at San Antonio de Ibarra; leatherwork at Cotacachi; woolen tapestries at Otavalo, Doctor Miguel Egas, and Salasaca; carpets at Guano; and Panama hats at Montecristi and near Cuenca. Folk music is equally rich, including the well-known *yumbo* and *el sanjuanito* from the highlands and the pasillo from the lowlands, as well as the varying local black and Indian (Amazonian, highland, and coastal) traditions. A revival of interest in folklore among the urban populations has led to the creation of folkloric dance troupes. Modern music is influenced by the Colombian *cumbia* and the Caribbean *salsa* and recorded by Ecuadoran groups with local themes.

Folk architecture is also rich, with varying traditions in bamboo, adobe, rammed earth, wattle and daub, and wood; modern architects have come to realize the continued potential of these traditions. The architectural monuments of the country include the large tolas (pre-Inca ramp mounds) of the northern highlands, such as those protected at the Cochasqui archaeological park; the Inca stone walls of Ingapirca near Cañar; the great colonial churches of Quito—especially San Francisco and la Compañía—with their paintings, statuary, and gilt wood carving; and the entire old urban centre of Quito, the site of a preservation and renovation project.

Modern fine arts are active, the best-known international figure probably being the painter Oswaldo Guayasamin, who is of mestizo-Indian parentage. A somewhat controversial figure, he earned an international reputation depicting the social ills of his society. Jorge Icaza's indigenist novel *Huasipungo*, which depicts the plight of Andean Indians in a feudal society, has also received international attention. Many novelists have come from the coast, including a "Guayaquil group" that explored life among the montuvio (mixed Indian, black, and white) population in a spirit of social realism; coastal novelists of note have included Luis Martínez, Demetrio Aguilera Malta, Joaquin Gallegos Lara, Enrique Gil Gilbert, Alfredo Pareja Diez-Canseco, and José de la Cuadra. Cuenca has been known for its poets, including Jorge Carrera Andrade and César Dávila Andrade. Books are published by both private and public presses, and the people have access to large book fairs and well-stocked bookstores.

Art and literature

Cultural institutions. The Central Bank of Ecuador, headquartered in Quito, sponsors some of the country's major historical and archaeological museums and research and also underwrites an active publishing program that produces *Cultura*, the country's premier cultural quarterly. The House of Ecuadoran Culture (founded in the early 1940s, with branches in many Ecuadoran cities) also sponsors cultural and historical research, publications, and special events; the National Historical Archives are a subdivision of this institution. The Ecuadoran Library "Aurelio Espinoza Polit," to the north of Quito in Chillogallo, is the country's premier library. The Central University Library, the National Library, the Pontifical Catholic University in Quito, and the Municipal Library in Guayaquil also have significant collections. Notable museums of archaeology and ethnology are located in Quito and Guayaquil.

Recreation. The Ecuadoran calendar is replete with religious and secular holidays. Some of the more important ones are not national but, rather, associated with local urban or regional traditions, such as the holidays of Quito (December 1–6), Guayaquil (October 9), and Cuenca (November 3) and the Yamor festival in Otavalo in early September. Many shops and businesses also close on Saturday afternoon and Sunday. Ecuadorans devote holiday periods to sports such as soccer (the national game), basketball, and volleyball and to picnics in the countryside, to excursions to the beach, or to socializing with family or friends. Cockfights are popular, and bullfights are occasionally held in the highlands. Glove ball is a highly popular attraction on Sunday afternoons in Quito and Ibarra. National and natural parks and preserves are relatively underused, although there is some interest in mountaineering and fishing. Many holidays are associated with particular foods or drinks, and music, live or recorded, is a part of most celebrations. Uniquely popular in Ecuador are beauty contests, which are held frequently at all levels of society throughout the country.

Press and broadcasting. Many Ecuadorans are avid readers, and they support numerous newspapers and periodicals. *El Comercio*, published in Quito, is perhaps the country's most prestigious newspaper; it provides detailed, serious coverage of political, economic, environmental, and cultural news, together with commentary by a number of well-known columnists. A wide range of points of view are expressed in other newspapers and periodicals; there is no censorship generally, but debate about the validity of Ecuador's territorial claims is strictly forbidden by the government. *Vistazo*, in Guayaquil, is the most popular magazine, covering national news events and personalities in a lively and often irreverent fashion. Radio stations include one of the oldest and most powerful transmitters

Newspapers and periodicals

in the Andes, La Voz de los Andes, which is affiliated with Evangelical Protestant missionaries but provides a diverse fare of programming. Other stations broadcast everything from international rock to local pasillos, Latin-American rhythms, and Quechua language, music, and news programs. Television stations broadcast a range of soap operas, game shows, and imported programs, along with special coverage from the United States, Venezuela, Mexico, Argentina, and elsewhere. (H.P.V./Gr.W.K.)

For statistical data on the land and people of Ecuador, see the *Britannica World Data* section in the BRITANNICA BOOK OF THE YEAR.

History

PRE-SPANISH ERA

The area presently known as Ecuador had a long history before the arrival of Europeans. Pottery figures have been discovered that date from 3000 to 2500 BC. The area was to some extent a frontier, exhibiting Colombian, Peruvian, and even Mexican influences. Even possible contacts with Japan have been suggested, but these are much debated. The territory, however, was also a distinct cultural area, inhabited by a variety of linguistic groups, including the highland Cara.

By AD 1400 Ecuador was divided into several warring states. In the early 15th century the Cara nation led by the Shyri dynasty began to expand in the northern and central Sierra. At approximately the same time both the larger Chimu nation of northern coastal Peru and the growing Inca state centred in Cuzco began to exert influence and pressure.

Inca conquest

The Inca conquest of Ecuador was begun by Topa Inca Yupanqui (ruled 1471–93) and extended by his successor, Huayna Capac (ruled 1493–1525), who lived much of his later life in Tomebamba, Ecuador. Although their cultural impact was otherwise spotty, the Inca spread the use of Quechua as a lingua franca and ordered large forced migrations where resistance to their conquest was especially strong. In Ecuador it is evident that Inca rule was resented by some and supported strongly by others. Huayna Capac left the Inca empire divided between his legitimate heir Huascar, in Cuzco, and his son by an Ecuadoran Cara princess, Atahuallpa. This led to a territorial dispute, and Atahuallpa won the ensuing civil war after a major battle near Riobamba in 1532; at just about the same time, a Spanish expedition led by Francisco Pizarro appeared off the coast. Atahuallpa was executed the next year as the Spanish conquest spread. In many parts of Ecuador Inca rule was less than 50 years old, and many of the pre-Inca states still held people's allegiances. As a result the Spanish under Pizarro's lieutenant Sebastian de Belalcázar were welcomed as liberators by some when they invaded Ecuador from Peru in 1534, while stiff resistance was encountered from others, especially the local leader, Rumiñahui, who was captured by the Spanish and executed in Quito.

THE COLONIAL PERIOD

Spanish estates in the Sierra

In the mountainous Andean area of central Ecuador (the Sierra), the Spaniards established a colony of large estates worked by Indian peons. People lived in semiautonomous Indian villages or in Spanish and mestizo administrative and religious centres such as Quito, Ambato, and Cuenca. The making of rough textiles in primitive sweatshops was the only industry.

On the Pacific coastal plain (the tropical Costa), there were fewer Indians to do the work, and the area was extremely unhealthy until the advent of modern medicine. As a result, the coast was neglected during the colonial period, although there was some shipbuilding and exporting of cocoa from the port of Guayaquil. The small coastal population of mixed races, with plenty of vacant land and less coercion of labour, developed a very different culture from that of the Sierra.

On the eastern slopes between the Andes and the headwaters of the Amazon (the Oriente), recalcitrant Indians and the equatorial climate prevented settlement, and the only Spaniards who attempted to live there in any numbers were missionaries. Later, this demographic vacuum was to cause Ecuador many problems.

The country's fourth major subdivision, the Galápagos Islands, were little more than pirate nests during the colonial period; but they were to achieve world fame in the 19th century because it was there that Charles Darwin made a major portion of the observations that led to his theories on evolution and the *Origin of Species*.

The people of Quito, the Ecuadoran capital, claim that it was the scene of the first Ecuadoran patriot uprising against Spanish rule (1809). Invading from Colombia in 1822, the armies of Simón Bolívar and Antonio José de Sucre came to the aid of Ecuadoran rebels, and on May 24 Sucre won the decisive Battle of Pichincha on a mountain slope near Quito, thus assuring Ecuadoran independence.

EARLY NATIONAL HISTORY, 1830–C. 1925

Ecuador's early history as a nation was a tormented one. For some eight years it formed, together with Colombia and Venezuela, the confederation of Gran Colombia. But in 1830, after a period of protracted regional rivalries, Ecuador seceded and became a separate, independent republic.

Sierra-Costa political rivalry

Liberal-Conservative hostilities. An increasing rivalry and ideological difference between the Sierra and the Costa usually focused on the two leading cities—Quito, the capital, in the Sierra, and Guayaquil, the country's principal port. Quito was the home of a landed aristocracy, whose positions of power during this early period were based on large, semifeudal estates worked by servile Indian labour; it was (and to some extent has remained) a conservative, clerical city, resistant to changes in the status quo. Guayaquil, on the other hand, by the 19th century had become a bustling, cosmopolitan port, controlled by a few wealthy merchants. These men and those around them were influenced by 19th-century liberalism; interested in trade, they favoured free enterprise and expanding markets, and some were anticlerical. Their bourgeois attitudes conflicted sharply with the more aristocratic beliefs of the Sierra elites. These early rivalries tended to be exacerbated by the nature of the two cities. The people of Guayaquil, the nation's breadwinner and the home of Ecuador's industry and trade, felt that a disproportionate part of the state's tax income was spent in Quito by government bureaucrats. Those in Quito complained that their exports had to pass through the monopolistic bottleneck of Guayaquil, which acted as a traditional middleman and, by adding to the price of Sierra products, reduced their competitiveness in the world market.

Rivalry between Flores and Rocafuerte (1830–45). Ambitious generals and politicians have played on this Quito-Guayaquil rivalry since the foundation of the republic in 1830. During the period 1830–45 two leaders from the wars of independence—Juan José Flores and Vicente Rocafuerte—struggled for power; Flores found much of his support in Quito, Rocafuerte in Guayaquil. Hostility was not constant, and for a few years the rivals agreed to alternate in the presidency. They were not simply personalist dictators; Rocafuerte in particular had a coherent ideology of government and did much to improve the educational institutions of the main cities. Both, however, were capable of deplorable conduct in their efforts to retain or regain power. Flores, on one occasion, even invited the Spaniards to return.

Breakdown of national government (1845–60). The rivalry between Flores and Rocafuerte was a struggle between two strong leaders. Between 1845 and 1860, however, the nation went through a period of chaos in which a series of squabbling, weak leaders (usually self-proclaimed liberals) fought for the presidency. This period reinforced the already close ties between the military and the national government.

The regime of García Moreno (1860–75). In the next period (1860–75) one of Latin America's most extraordinary experiments in autocracy occurred, during the presidency of Gabriel García Moreno. As a young man García Moreno had witnessed the chaos in Ecuador and the selfish struggles of the various cliques. He had also seen the European Revolutions of 1848 and had developed an

abhorrence of liberalism and of uncontrolled violence. A careful analysis of Ecuadoran society led him to conclude that the young nation lacked unifying factors; it had no great tradition, suffered from regional resentments, and was sharply divided by class and between whites and Indians who did not even share a common language. García Moreno concluded that the only social "cement" was religion—the general adherence of the population to the Roman Catholic church. He felt that in time nationalism could be created and that more social cohesion would emerge as a result, but that meanwhile Ecuador needed a period of peace and strong government. When he became president, therefore, he based his regime on two factors—strong authoritarian personal rule and the Roman Catholic church. All education and welfare, and the direction of much government policy, were turned over to clerics. Other religions were harshly discouraged. All opposition was ruthlessly suppressed, and some leading liberals spent many years in exile.

Lack of unifying elements

Although many aspects of García Moreno's regime were reactive, it did mark the first period of genuine progress for Ecuador. Roads, schools, and hospitals were built. A start was made on a Quito-Guayaquil railroad, to tie together the Costa and the Sierra. García Moreno encouraged the planting of eucalyptus trees from Australia to combat erosion in the Sierra, where the original ground cover had been cut down for fuel by the impoverished Indians. Other agricultural reforms slowly raised production. By the end of his regime a strong feeling of nationalism had been created among the urban classes.

In the 19th century, however, this authoritarian, clerical government seemed an anachronism, and liberal opposition grew both at home and abroad. When García Moreno was assassinated on the steps of the government palace in 1875, the liberal intellectual and pamphleteer Juan Montalvo proclaimed from exile, "My pen has killed him."

Shift to liberalism (1875–97). García Moreno's death, as he himself might have forecast, brought a period of near anarchy. Conservatives and Liberals struggled for power. But Ecuador had become part of the world market; the importance of the coast slowly increased, and the Liberals of that area more and more dominated the economy.

A new Liberal hero emerged from the lower classes as the leader of the coastal reaction to Sierra conservatism and clericalism. A man of great personal magnetism, General Eloy Alfaro, led a march against the Sierra in 1895 and after a year became constitutional president, serving two terms (1897–1901 and 1906–11). Much of the administrative structure of the García Moreno era was dismantled. The anticlerical Liberals gradually removed the church from state education: they instituted civil marriage and burial, proclaimed freedom of religion, permitted divorce, and eased controls on the press. The church's tithe was abolished, and many of its large estates were confiscated by the state, some estates passing into the hands of Liberal leaders.

The reforms of Eloy Alfaro

In many ways, however, in spite of political manifestos to the contrary, the Liberals of this era shared the basic ideas of the previous period. They advanced García Moreno's road- and railroad-building programs; the Quito-Guayaquil railroad was completed in 1908 during Alfaro's second term. Moreover, central government did not lose its authoritarian caste; Alfaro, the Liberal strongman (caudillo), was as arbitrary and ruthless as his conservative predecessor. In the Sierra and on the coast, power remained unchanged. The problem of the great haciendas was not touched, and the change to liberalism meant little to the impoverished Indians and peasants.

Alfaro's overthrow, like that of García Moreno, was brought about by his stubborn attempts to perpetuate himself in office. A coalition of Conservatives and dissident Liberals forced him and his clique from the presidency in August 1911, but when the next president died in office shortly thereafter, the aging and increasingly unpopular Alfaro returned from exile and tried to recapture his following. The leaders of the Liberals rejected him, and after some fighting he was arrested in Guayaquil. He and his lieutenants were sent to a model prison in Quito, built years before by García Moreno. There, on Jan. 28, 1912, a lynch mob broke in, dragged the prisoners through the streets, and burned the bodies.

Problems of the early 20th century. The Liberals remained in office, but the real power continued to rest in the hands of the wealthy merchants and bankers of Guayaquil. During World War I and the short boom that followed it, this clique further extended its influence and diversified its capital with a view to owning the agriculture of the coastal plain. Cacao was the dominant export crop, as in the colonial period, but sugar and rice became increasingly important.

A depression followed in the early 1920s. The price of food increased, and exports in general declined. The sucre—the national unit of currency—fell rapidly in value. At the same time, the nation's cacao plantations became infected with a fungus known as witches'-broom, and production sagged. These crises brought urban discontent, the formation of trade unions in Guayaquil, riots, and massacres by the army. Hundreds died during riots and shootings in November 1922.

In 1925 the army entered this turbulent situation, claiming that it wished to restore national unity and blaming many of the country's problems on the merchant bankers of Guayaquil. Like most Latin-American revolutions, that of 1925 brought little change in social and economic structures.

MODERN HISTORY

The period between 1925 and 1948 was one of greater turbulence than Ecuador had ever known. Increasing involvement in the world market and in international politics meant that the nation could no longer escape entanglements and the consequences of world ideological conflicts. Yet during this crucial period, Ecuador's internal disunity prevented the modernization of its social structure, land tenure system, education, and communications. Thus, the country was badly equipped to face the demands of the age.

Economic development and loss of territory in the 1940s. Ecuador was still suffering from the effects of the Great Depression when it became involved in World War II. It sided with the Allies and allowed the United States to build military bases on its territory, but it played little direct part in the war. Under President Carlos Arroyo del Río, Ecuador drew some benefit from the higher prices for raw materials caused by the war, and the early years of the war were relatively prosperous and tranquil.

World War II had a serious secondary effect on the nation, however. Because of lack of capital and people, Ecuador by 1940 had not effectively settled its vast Amazonian territories. In July 1941, after long diplomatic bickering and a series of border incidents, the Peruvian army invaded, seized much of the disputed Amazonian area, and devastated El Oro *provincia* (Ecuador had lost territory to each of its more powerful neighbours during its troubled history). The Ecuadoran forces, poorly trained and equipped, were easily defeated, and the disgrace caused the overthrow of Arroyo del Río. The United States and the other major powers were too preoccupied with the war to allow such small conflicts to destroy Allied unity or to disrupt the production of vital raw materials. A peace conference in Rio de Janeiro in 1942 forced Ecuador to relinquish its claims to much of the Amazonian region. Subsequently, Ecuador repeatedly attempted to reopen the question, claiming that the Protocol of Rio was imposed by force and that the new borders were therefore invalid. This constant Ecuadoran irredentism was used repeatedly by demagogues and ultranationalists, who distracted attention and effort from urgent internal problems.

War with Peru

Domination of Velasco Ibarra (post-World War II). Politics and government after World War II presented contradictions. Ecuador enjoyed a long period of constitutional government and relatively free elections following the presidency of the Radical Liberal leader Galo Plaza (1948–52). There were also two long interludes of military government (1963–66; 1972–79), but the period was dominated by one of Latin America's great caudillos, José María Velasco Ibarra. Velasco Ibarra, who died in 1979, was president of Ecuador five times but completed only

one of these terms. He seemed able to win any election, such was his popularity with the masses, but his terms of office were marked by sudden reversals in policy, contradictory economic programs, personal outbursts, temporary suspensions of civil liberties, and military interventions. Some critics claimed that Velasco Ibarra drew support from communist groups; others said he was the puppet of powerful business groups in Guayaquil. But no group was able to control the erratic leader for long.

Impact of Velasco Ibarra's rule

Velasco Ibarra's political presence may have inhibited the development of social and economic reforms. His personal appeal cut across parties and ideologies. The traditional parties—the Liberals and the Conservatives—were thrown into disarray by his incursions, and the growth of newer parties such as the Ecuadoran Socialists and the Social Christians was retarded. Opponents alleged that Velasco Ibarra made economic progress impossible because he so frequently halted or reversed measures undertaken by previous governments.

Ecuador from the late 20th century. After Velasco Ibarra's last fall from power, in 1972, military officers ruled for some seven years before handing over the government to a constitutionally elected (July 16, 1979) civilian president. The civilian and military governments of the 1970s did not adequately manage the oil boom that occurred in that decade. The boom increased the size and wealth of the middle class, led to the building of roads, quays, pipelines, and other infrastructural features, and caused severe inflation. The poor suffered the effects of inflation but reaped few of the benefits of the oil boom.

Velasco Ibarra's death and the withdrawal of the military officers from government allowed the nation to return to an elected civilian government and a new constitution in 1979. Jaime Roldós Aguilera, a young social democrat, was elected president on a reformist platform. He promised greater social equality and a more equitable distribution of oil industry profits, but he was unable to manage the legislature and was soon at odds with his own party. His popularity increased after a border skirmish with Peru in early 1981, but he was killed in an airplane crash later that year. His successor was Osvaldo Hurtado Larrea of the small Christian Democratic party. The economy, depressed by a drop in world oil prices, spiraled downward with accompanying high inflation and a depreciating currency. León Febres Cordero, a congressman from Guayaquil, was elected president in 1984. His free-market economics and pro-U.S. foreign policy drew Ecuador into closer alliance with the U.S. administration of President Ronald Reagan, but Febres Cordero was never popular in Ecuador. Oil prices continued to fall, and his troubles with the National Congress and the military led to calls for his resignation and, on one occasion, to his being kidnapped by air force personnel for half a day until he agreed to release one of their leaders. In March 1987 he suspended interest payments on Ecuador's $8.3 billion foreign debt after an earthquake destroyed part of a major oil pipeline.

In the presidential runoff election of May 1988, a left-wing opponent, Rodrigo Borja Cevallos, was elected, but he seemed to have few solutions to the steadily worsening economic crisis. (M.J.MacL.)

Indigenous uprising of 1990

Borja's term was marked by a major indigenous uprising in June 1990, which was remarkably nonviolent. Thousands of Indians marched, blockaded roads, and occupied churches as they demanded land reform, recognition of Quichua (Quechua) as an official language, payment by oil companies for damage to indigenous lands, and support for bilingual education. The Ecuadoran government subsequently recognized indigenous land claims to much of the Oriente (but not to subsurface resources) and further promoted bilingual education.

The moderate-conservative Sixto Durán Ballén was elected president in the runoff election of July 1992. Durán eliminated most subsidies, balanced the government's budget, reduced trade barriers, brought Ecuador into the World Trade Organization, and encouraged foreign investment. These measures helped to reduce inflation, but interest rates remained high and the economy grew only slowly. A border war with Peru in early 1995 led to a standoff and proved that the Ecuadoran military had improved its capabilities; however, it left the country with a crippling war debt. Also that year, the vice president fled the country to avoid arrest on corruption charges.

Power swung to the coast in the July 1996 election when the populist leader Abdalá Bucarám Ortiz was elected president. Bucarám raised prices and became increasingly unpopular because of his seemingly erratic and controversial behaviour. In February 1997, the Congress forced Bucarám into exile and replaced him with Fabián Alarcón Rivera; voters endorsed that action in May. In early 1997 torrential El Niño rains caused major damage to the coast and further weakened the ailing Ecuadoran economy. The July 1998 elections placed Jamil Mahuad Witt, the mayor of Quito, in the presidency. In order to fight inflation, he imposed austerity measures, which were countered by public protests and demonstrations. Mahuad and President Alberto Fujimori of Peru agreed to a comprehensive delineation of the countries' long-disputed frontier; the last of the new border markers was placed in May 1999.

Dollarization

The nation's economy stagnated and its currency remained unstable. After several banks failed, Mahuad announced a plan to designate the U.S. dollar as the nation's new currency. Many groups of students, labourers, and Indians opposed the plan. Civil unrest intensified until January 2000, when a military coup deposed Mahuad and elevated Vice President Gustavo Noboa to the presidency. However, Noboa announced few initial changes to the government's policies, and the U.S. dollar became official Ecuadoran currency in September. Thousands of Ecuadorans, pressed by economic hardship, emigrated to Spain, other European nations, and the United States during the early years of the 21st century. During the same period Ecuadorans felt increasingly threatened by the escalating violence in neighbouring Colombia. (Ed.)

For later developments in the history of Ecuador, see the BRITANNICA BOOK OF THE YEAR.

For coverage of related topics in the *Macropædia* and *Micropædia*, see the *Propædia*, sections 966 and 974, and the *Index*.

BIBLIOGRAPHY

Physical and human geography: The many volumes of CENTRO ECUATORIANO DE INVESTIGACIÓN GEOGRÁFICA, *Geografía básica del Ecuador* (1983–), provide useful overviews. A good thematic atlas is LOUIS ARRÉGHINI and JUAN LEÓN (eds.), *Ecuador: espacio y sociedad* (1997). The country's physical geography is examined by ALAIN WINCKEL, CLAUDE ZEBROWSKI, and MICHEL SOURDAT, *Las regiones y paisajes del Ecuador* (1997); and ANNE COLLIN DELAVAUD (ed.), *Atlas del Ecuador* (1982). Discussions of plant and animal life include ERWIN PATZELT, *Fauna del Ecuador* (1989); and ROLF WESCHE, *The Ecotourist's Guide to the Ecuadorian Amazon: Napo Province* (1995).

Ethnographies of various Indian groups include NORMAN E. WHITTEN, JR., *Sacha Runa: Ethnicity and Adaptation of Ecuadorian Jungle Quichua* (1976); and PETER BROENNIMANN, *Auca on the Cononaco: Indians of the Ecuadorian Rain Forest* (1981). MARY J. WEISMANTEL, *Food, Gender, and Poverty in the Ecuadorian Andes* (1988), focuses on the central highlands. GREGORY KNAPP, *Geografía quichua de la sierra del Ecuador* (1987), analyzes ethnicity, whereas his *Ecología cultural prehispánica del Ecuador* (1988) focuses on traditional highland agriculture.

Economic and developmental studies include DAVID W. SCHODT, *Ecuador: An Andean Enigma* (1987); MARCO A. ENCALADA REYES, *Medio ambiente y desarrollo en el Ecuador* (1983); and *Ecuador: An Agenda for Recovery and Sustained Growth* (1985), an assessment by the World Bank. JEAN PAUL DELER, *Genèse de l'espace équatorien* (1981), also available in a Spanish trans., *Ecuador: del espacio al estado nacional* (1987), is a study of spatial organization. JUDITH KIMERLING and SUSAN HENRIKSEN, *Amazon Crude* (1991), is a lively discussion of the politics of oil and Indians in the Oriente.

History: ALFREDO PAREJA Y DÍEZ CANSECO, *Historia del Ecuador*, 2nd ed. rev., 2 vol. (1958), is a standard history. PABLO CUVI, *Velasco Ibarra* (1977), contains essays on the great caudillo and his times. Events of the 1900s are traced in FRANK MACDONALD SPINDLER, *Nineteenth Century Ecuador: A Historical Introduction* (1987). OSVALDO HURTADO, *Political Power in Ecuador* (1981; originally published in Spanish, 1977), is a perceptive historical analysis of economic and political power. DAVID CORKILL and DAVID CUBITT, *Ecuador: Fragile Democracy* (1988), studies recent politics and problems. ENRIQUE AYALA MORA, *Nueva historia del Ecuador*, 15 vol. (1983–95), is a multifaceted collection of historical essays.

(Gr.W.K./M.J.MacL./Ed.)

Edinburgh

Edinburgh (Gaelic: Dun Eideann), the capital city of Scotland, is centred near the southern shore of the Firth of Forth, an arm of the North Sea that thrusts westward into the Scottish Lowlands. The city and its immediate surroundings constitute an independent council area. The city, including the port of Leith on the Firth of Forth, lies within the historic county of Midlothian.

Physically, Edinburgh is a city of sombre theatricality, with much of this quality deriving from its setting among crags and hills and from its tall buildings and spires of dark stone. Edinburgh has been a military stronghold, the capital of an independent nation, and a centre of intellectual activity. Although it has repeatedly experienced the vicissitudes of fortune, the city has always renewed itself. Today it is the seat of the Scottish Parliament and the Scottish Executive. It remains a major centre for finance, law, tourism, education, and cultural affairs.

This article is divided into the following sections:

Physical and human geography 963
- Character of the city 963
- The landscape 963
  - Climate
  - The city layout
  - The Old Town
  - Holyrood
  - Princes Street Gardens
  - The New Town
- The people 965
- The economy 965
  - Industry
  - Commerce and finance
  - Transportation
- Administration and social conditions 966
  - Government
  - Health
  - Education
- Cultural life 966

History 966
- The early period 966
- The medieval city 966
- Union with England 967
- The modern city 968

Bibliography 968

Physical and human geography

CHARACTER OF THE CITY

Although Edinburgh absorbed surrounding villages and the Firth of Forth ports between 1856 and 1920, its mood is still generated within the small compass of the Old Town and the New Town. The Old Town, built up in the Middle Ages when the fear of attack was constant, is huddled high on the Castle Rock overlooking the surrounding plain; the New Town, in contrast, spreads out in a magnificent succession of streets, crescents, and terraces.

"This profusion of eccentricities, this dream in masonry and living rock is not a drop-scene in a theatre," wrote Robert Louis Stevenson, the 19th-century Scottish novelist, essayist, and poet, who was born in the New Town, "but a city in the world of reality." The contrasts that make Edinburgh unique also make it typically Scottish, for, despite its reserved exterior, it is also a city capable of great warmth and, upon occasion, gaiety. Its history demonstrates that its citizens have also been capable of great passion, especially in matters royal or religious. In 1736 the burgh almost lost its royal charter following the lynching of one John Porteous, captain of the city guard. This occurrence can be viewed as a violent gesture common to the history of most old cities; yet, even in this moment of conventional madness, the city manifested its complex character: needing a hanging rope, the mob descended on a shop and bought one.

A city long renowned for a somewhat inflexible respectability—when West Princes Street Gardens were turned over to the general public in 1876, smoking was forbidden—Edinburgh concurrently maintained a fascinating netherworld of ribaldry and drunkenness. A poet or jurist of sufficient distinction might succeed in inhabiting both worlds. Such "Edinburgh characters" abounded during the flourishing neoclassical period of the 18th and 19th centuries known as the Augustan age, when the city's authors, critics, publishers, teachers, physicians, and scientists formed an intellectual elite of world influence.

(B.E./A.R.T.)

THE LANDSCAPE

Until late in the 18th century Edinburgh followed a common European pattern in continually renewing itself on its original site, but the lack of space for outward expansion compelled each successive phase to conform to the original layout. Subsequently, when expansion became possible, almost at a bound the town broke free of its medieval mold, and each further refashioning was built adjacent to, rather than on top of, its predecessor. In consequence, the soaring vertical lines of the Old Town confront the expansive horizontal ones of the Georgian New Town, and both are encircled by acres of individually distinct Victorian suburbs and finally by a ring of 20th-century urban sprawl that makes its way toward hills and sea.

Climate. Edinburgh has a mild climate. Its proximity to the sea mitigates temperature extremes. Winters are relatively warm, with average daily minimum temperatures remaining above freezing; and summers are comparatively cool, with temperatures seldom rising much above 70° F (21° C). The prevailing easterly winds are often cold but relatively dry; warmer southerly winds coming off the North Atlantic Current often bring rain. Annual precipitation is moderate, averaging 27 inches (676 millimetres), and it is evenly distributed throughout the year. Edinburgh lacks prolonged sunshine: on average for the year it receives less than one-third of the possible sunshine for its latitude. But even this is in part compensated for by its ever-changing cloudscape.

The city layout. Edinburgh now occupies some five miles (eight kilometres) of north-facing slope between the Pentland Hills and the broad estuary of the Forth River, merging there with the once-independent seaport of Leith. This slope is punctuated by upthrusts of lava; one such, called Arthur's Seat, rises some 820 feet (250 metres) and dominates the southeastern flank. The valleys between these striking hills were scoured deep and clean by glacial action in the Pleistocene Epoch. Edinburgh has been built on top of and around these obstacles so that the nearer one comes to the city centre, the more spectacular is the juxtaposition of natural and built environment, with terraces of stone confronting soaring thrust.

Castle Rock

At the city's core is the Old Town's Castle Rock, a plug of black basalt sealing the vent of an extinct volcano. It rises 250 feet from the valley floor and is crowned by the famous castle, which, subtly floodlit every night, stirs even the habituated townsfolk. Glacial ice once flowed from the west and around the Castle Rock's flanks, depositing to the east of the rock the accumulated debris of a lateral moraine—called a crag and tail formation. Along the crest

of this tail, and down its steep sides, the Old Town was built from the 11th century onward. The shortage of building space compelled it to expand skyward, and 10- and even 12-story tenements separated by narrow passageways were built in the 16th and 17th centuries.

Two hundred yards (183 metres) north of the Castle Rock, across the valley that is now Princes Street Gardens, lies the New Town, a district that was planned and built in successive phases between 1767 and 1833. It offers a dignified tribute to the international taste of the Enlightenment and to the surveyor's set square; its design was overly regular to begin with, but later developments paid more respect to natural contours and softened the regimentation of the right angle with curves and crescents. The New Town's northwestern boundary is, roughly, the line of Edinburgh's only stream, the Water of Leith. The stream's brief course from the Pentlands to the sea provided power for the mills of a series of villages—Dalry, Dean, Stockbridge, Silvermills, and Canonmills—that grew up beginning in the 16th century. The villages are now embedded in the 19th-century matrix of the town.

Expansion from the Old Town. For centuries the barrier to northward expansion was the lake and encircling marsh—the North Loch, or Nor' Loch—that choked the valley along the foot of the moraine and the Castle Rock. The lake had been created from swampland by James II as a defense against attack. Even when it was drained and the land was firmed, access to the north had to await the ability of civil engineers to span the valley with a bridge. This was achieved in 1772, when the North Bridge, 70 feet high and 1,130 feet in length and canted steeply northward, was completed.

In the centuries between the founding of the Old Town and the beginning of the New Town, Edinburgh eased itself down the southern flank of the moraine. A marketplace, known as the Grassmarket, was built in the shadow of the Castle Rock, and a roadway known as Southgate (later as Cowgate) led eastward to the Augustinian Abbey of Holyrood (later the site of the Palace of Holyroodhouse). The town then slowly climbed the facing slope of the adjacent moorland, despite the dangers inherent in exposing a southern flank to English assault. Well before the North Bridge was built, this southern suburb had grown up, dignified with fine churches—Magdalen Chapel (1544), Greyfriars (1612)—and the "Toun's College" (later the University of Edinburgh). The suburb was composed in a small-scale domestic architecture that is noticeably friendlier than the somewhat bleak facades of the New Town.

Edinburgh's bridges. In the 50 years following the building of North Bridge, four other bridges were completed, enabling the city to expand where it pleased. Two of these, South Bridge (1788) and King George IV Bridge (1834), are multiple-arch constructions that span the Cowgate ravine. These opened the way south to rapid expansion. In the same period Waterloo Bridge, with its Regency Arch (1820), opened the eastern slopes of Calton Hill (northeast of the Castle Rock) to Regency building, while King's Bridge (1833), leaping westward from the Castle Rock, was the vital link in the so-called "western approach." Throughout the Victorian and Edwardian ages the city grew in every direction, recording in its stone tenements and detached mansions every foible of changing taste—Neoclassical, Gothic, Scotch Baronial, Italianate—to spend itself in 20th-century brick and concrete. (A.R.T.)

Edinburgh Castle

The Old Town. Edinburgh Castle, 443 feet above sea level, dominates the city. The Castle Rock may have been fortified as early as the 6th century AD, but continual use of this strategic site has obliterated all material evidence of occupation before the 11th century. The principal surviving buildings have been altered to suit changing needs. They are of greater historical than architectural interest, and they date mainly from the 16th century. The small chapel of St. Margaret, the queen of Malcolm III Canmore, on the highest point of the rock, dates from the 12th century and is the oldest surviving building.

The Royal Mile

The Royal Mile, which begins outside the Castle Esplanade, descends Castle Hill, the crest of rock linking the castle with the Palace of Holyroodhouse to the east. The Royal Mile bears several names—Castle Hill, Lawnmarket, High Street, and Canongate—recalling its medieval districts. The Old Town's towering tenements are pressed together along the crest, and the cliff-face of houses is broken by wynds—narrow, winding, stone lanes leading down either side of the ridge—and closes or vennels—entryways into courtyards, around and behind which are yet more buildings.

The High Kirk

A number of important buildings are located along the Royal Mile. At its heart is the High Kirk, or Cathedral, of St. Giles, the National Church of Scotland. It has a fine late Gothic nave and is surmounted by a magnificent 15th-century crown tower, an open spire with eight flying buttresses supporting a sculptured turret. Much of the exterior stonework is a 19th-century restoration.

Behind St. Giles, in Parliament Square, is Parliament House, which was built by the Town Council during 1632–39 with the aid of private contributions when Charles I threatened to move the law courts elsewhere. The Scots Parliament met there from 1639 to 1707. The supreme civil and criminal courts of Scotland now sit in the building. The Parliament House complex is adjacent to the National Library of Scotland, successor to the earlier Advocates' Library, of which the philosopher David Hume was once librarian.

Opposite St. Giles, slightly to the east, is the City Chambers. The building was completed in 1761 as the Royal Exchange but never was used for that purpose. It faces the Mercat Cross (Market Cross), the hub of the old city. Part of the original pillar of the Cross was used in a late 19th-century reconstruction undertaken at the expense of Prime Minister William Gladstone. Royal proclamations are announced from its tower platform by members of the Court of Lord Lyon (the Scottish equivalent of the College of Heralds).

Almost opposite the Tron Kirk, where all incoming goods once passed over an iron scale ("tron"), and where the North and South bridges cut across the crest, is Anchor Close, where William Smellie, printer to the University of Edinburgh and editor of the first edition of the *Encyclopædia Britannica,* printed the 1787 edition of the *Poems* of Robert Burns. Smellie introduced Burns to the conviviality of a club called the Crochallan Fencibles; the poet, in return, regaled them with the bawdy songs entitled *Merry Muses of Caledonia* (first published in 1800).

Noteworthy buildings along the High Street and Canongate sections of the Royal Mile include the John Knox House, built in the late 15th or early 16th century; Moray House, a 17th-century town house now used as a teacher-training college; the baroque-fronted Canongate Church (1688–90), in the graveyard of which are the tombs of Robert Fergusson, an 18th-century poet, and the political economist Adam Smith; Acheson House (1633), containing the Scottish Craft Centre; Huntly House (containing the Civic Museum); and the old Canongate Tolbooth (1591). The 17th-century inn in White Horse Close, once the terminus of the London coach, has been splendidly restored as private homes.

Holyrood. Away from the crowded buildings, at the lower end of the Royal Mile, is Holyrood. The abbey was founded in 1128 and rebuilt about 1220. The ruins of the nave are impressive examples of the bold, imaginative work of the period. The Palace of Holyroodhouse, which shouldered the abbey aside, comprises an early 16th-century wing and a 17th-century quadrangle court and facing wing. It is the sovereign's official residence in Scotland. South and east of the palace is Holyrood Park, which includes the Salisbury Crags and Arthur's Seat as well as three lochs.

The floral clock

Princes Street Gardens. The Princes Street Gardens, laid out between the Old and New towns in the drained lakebed of the old North Loch, have a distinct character. Flowers are set out in beds changed several times a year, and a floral clock planted in 1903 (the first in the world), which embowers a quarter-hour cuckoo, has some 24,000 plants in its 36-foot circumference. Among the lawns, flower beds, and groves are recreational areas, a bandstand, an outdoor dance floor, and numerous memorials, the most conspicuous of which is an 1844 Gothic spire,

200 feet high, that rises above a statue of Sir Walter Scott and his hound, Maida.

For the first 100 years of its existence, West Princes Street Gardens was the private amenity of Princes Street proprietors. In 1876 this tract was opened to the public, which had always had access to the eastern gardens. The Mound, a causeway of rubble and earth from New Town construction, forms the division between the two gardens. On the Mound stand two neo-Grecian temples to the arts: the Royal Scottish Academy (1832) and the National Gallery of Scotland (1859). Railroad tracks—now almost concealed by landscaping—were laid through the middle of the gardens in 1847, and trains bound for Glasgow and the north pass under the Mound. The rails terminate in the east end of the park at Waverley Station.

The New Town. The town council approved plans for the New Town in 1767. It was designed to be a residential district. The architect, James Craig, set out a grid five streets deep and seven streets wide, its broad central axis terminating in grand squares at each end. St. George's Church would anchor the western end of the scheme, St. Andrew's the eastern. Princes Street, the southernmost of the new streets, was lined only on its north side with residences, which faced the castle across the valley. The arrival of the railway changed the face of Princes Street as residential space gave way to shops and hotels. Princes Street became the main shopping street and the principal thoroughfare of the city, and few original buildings remain behind the shop fronts. The Register House (1772–92), at the east end of Princes Street facing the North Bridge, is the finest of the city's buildings by the 18th-century architects James and Robert Adam. It houses part of the national records of Scotland. In the remainder of Craig's New Town much has been done to restore and improve the amenity of the streets and squares. In St. Andrew Square the Royal Bank of Scotland, built as a town house in 1772–74 for Sir Lawrence Dundas when he was the member of Parliament for Edinburgh, is a fine example of an 18th-century mansion and has a stunning Victorian banking hall (1858). In George Street is the parish church of St. Andrew, an oval building with a fine plaster ceiling and an elegant spire. On the north side of Charlotte Square the Georgian House, managed as a museum by the National Trust for Scotland, is completely furnished, from kitchen to bedrooms, with all the appurtenances of late 18th-century Edinburgh elegance.

At the east end of Princes Street the Calton Hill rises above the central government office of St. Andrew's House (1937) and the adjacent Royal High School (1825–29). It is crowned by the 19th-century architect William Playfair's City Observatory (1818) and a charming Gothic house, by Craig, built for the astronomer royal. Behind this rise 12 columns of an intended replica of the Parthenon, designed by Playfair in 1822 as a memorial to the Scots who died in the Napoleonic wars. Construction of the memorial was abandoned when the subscription failed in 1830. Down the slope to the south stands the tiered circular tower of the Nelson Monument (1807).

To the north, on the flat plain toward the Forth, the Royal Botanic Garden, at its finest when the great rhododendrons are in bloom, offers from its crest a superlative vista of the New Town backed by the distinctive skyline of the Old Town.

THE PEOPLE

Edinburgh's population is largely a mixture of middle-class professionals and nonprofessionals. Both have achieved a measure of prosperity as compared with their ancestors, most of whom emigrated from the surrounding countryside and small towns to provide the 19th-century city with unskilled and semiskilled labour. The work force has changed considerably since then, and now white-collar workers equal blue-collar workers in number. The city has only a small percentage of people from other countries; according to the 1981 census, less than 5 percent of the city's population was born outside of the United Kingdom.

THE ECONOMY

Industry. Edinburgh is today, as it was in the 18th century, predominantly a provider of services. Barely one-tenth of its residents are now in manufacturing. The pre-World War II staples of brewing, baking, and book printing have all suffered. Electrical and electronic research and engineering, much of it related to defense and much of it drawing on the scientific skills of the town's two universities, has become the largest industrial employer.

The main service industries are related to public administration, the law, medicine, investment and finance, education, and tourism: Edinburgh is second only to London as a British tourist city. The regional and district councils of local government constitute the largest employer, with education also of major importance.

The port of Leith, about 30 miles from the open sea, became a part of the six-port Forth Ports Authority in 1968 and was extensively modernized in the 1970s. Grain, foodstuffs, and wood products are the principal incoming products; outbound shipments consist of coal, whisky, iron, and steel. Granton, Edinburgh's other port town, is also under the authority, with special facilities for rapid loading and unloading; a sizable industrial estate to service North Sea petroleum has been developed there, and it is the home port for the Firth of Forth fishing fleet.

Commerce and finance. Edinburgh is an important centre for financial and legal services. The city's institutions financed much of the development of the American West, including ranching, railroads, timber, and mining, and thereby laid the basis for its fortune. As the centre of Scotland's legal system, Edinburgh has a flourishing legal

A.G. Ingram Ltd., Edinburgh



Princes Street, showing Castle Rock overlooking the city, the Scott Monument (foreground), the National Gallery of Scotland, and the Royal Scottish Academy, Edinburgh.

profession, which ranks second only to banking as the highest paid profession in the city.

Transportation. The city is served by ScotRail, the regional rail carrier. There is frequent train service to London and the major Scottish cities as well as regular service to other parts of Scotland and England. Edinburgh has two central railway stations: Waverley (the second largest in Britain) and Haymarket. Edinburgh's airport has no direct transatlantic service but does offer international service to continental Europe. The city has no subway system but has excellent bus service.

ADMINISTRATION AND SOCIAL CONDITIONS

Government. The Scottish Parliament is responsible for legislation concerning health, education, housing, economic development, regional transport, the environment, and agriculture. The leading parliamentary party elects a first minister, who heads the Scottish Executive, which implements Scottish legislation. Directly below the Scottish government is the popularly elected City of Edinburgh Council, which implements Scottish laws at the local level. It oversees services such as local planning, education, social services, housing, roadways and traffic, fire fighting, sanitation, housing, parks and recreation, libraries, city museums, and elections.

A decline in social service funding has been partly compensated for by volunteer organizations. Relatively high youth unemployment, as well as a lack of entertainment or recreational amenities (especially in some post-World War II suburbs), has made the city notorious for drug abuse, petty violence, and more serious crime.

Health. Even before the foundation of the Edinburgh faculty of medicine in 1726, the healing arts were both practiced and taught in the city. With the opening of the great new Royal Infirmary in 1748, however, Edinburgh became one of the chief medical centres in the world. Today the city has more than 10 hospitals. Edinburgh's medical community offers a range of health services unsurpassed anywhere in the United Kingdom.

An international medical centre

Education. The City of Edinburgh maintains a system of state schools that provide free primary and secondary education. The city also provides free nursery schools and schools for children with special needs as well as a program of community education for youth and adults. Edinburgh also has several fee-paying independent schools.

The University of Edinburgh, founded in 1583, is the city's largest university. A world-renowned intellectual centre for much of its history, it offers a range of undergraduate, postgraduate, and professional programs. Heriot-Watt University, dating from the earliest days of the Industrial Revolution, was one of the first of Britain's new technological universities. Much of its operation has been transferred to a satellite campus outside the city centre at Riccarton. Jewel and Esk Valley College offers a range of postsecondary vocational courses. The Edinburgh College of Art offers courses in the fine arts and various aspects of environmental design, including architecture, landscape architecture, and city and regional planning.

CULTURAL LIFE

Since 1947 Edinburgh has been an international focal point for the arts during the three weeks of its annual summer festival. The original Edinburgh International Festival now occurs concurrently with both the Edinburgh Festival Fringe, which features avant-garde and innovative performances, and the Edinburgh International Film Festival. Hundreds of thousands of visitors come for the theatre, ballet, music, films, art expositions, and the general excitement, which closes with a skirl of the Scottish bagpipes before the castle gate. A major cultural institution is the National Galleries of Scotland. It includes the National Gallery, with a fine international collection of art as well as a representative collection of Scottish painters. Under the direction of the National Galleries are the Scottish National Portrait Gallery and the Scottish National Gallery of Modern Art, which has a good collection of French Impressionist works.

The city has a large number of recreational facilities. In addition to spectator sport venues, it has within its boundaries 9 miles (14 kilometres) of coastline for boating, several beaches, and numerous golf courses. Indeed, golf has been played in Edinburgh since 1457. There are also scores of bowling greens, public association football (soccer) and hockey fields, cricket pitches, tennis courts, putting greens, and short-hole golf courses. The Meadowbank Sports Centre, just east of the city centre, has facilities for more than 30 sports. The Royal Commonwealth Pool is one of the finest in the United Kingdom. The city is home to the Hibernian and the Heart of Midlothian football clubs.

Recreation

(B.E./A.R.T.)

History

THE EARLY PERIOD

Settlement of the region. To the first settlers in Scotland, arriving in the early postglacial period as early as 7000 BC, estuaries and rivers offered the best access to the interior; and of these the Forth was among the most important. Its shoreline and mudflats show evidence of Stone Age explorers, who as yet had no need for the protection of the region's steep hills. Finds of swords and other metal objects suggest, however, that by about 1500 BC these Celtic tribes were making use of the crags of Arthur's Seat for defense. In the Iron Age, which in Scotland began in about 800 BC, hill forts proliferated in the Lothians—the area in the immediate vicinity of Edinburgh—and the Borders, to the south. There is no proof that the Castle Rock was occupied at that time, but it is unlikely that so good a defensive site would have been ignored. Holyrood Park, Blackford Hill, and Craiglockhart Hill, however, all bear witness to occupation in the late 1st millennium BC.

Strategic importance. The Romans viewed the Edinburgh plain between the Pentland Hills and the Forth as being of strategic rather than defensive value. During three or four decades in the second half of the 2nd century AD, the Antonine Wall, stretching across Scotland between the Clyde River and the Firth of Forth, was the northernmost defense in Roman Britain, and the site of Cramond, a village on the Forth within the modern city boundary, was the point at which one of Roman Britain's major north–south roads terminated. The road, with major forts at Dalkeith and Inveresk on the southeastern approaches to the present city, cut through what is now the Meadows district of Edinburgh and guarded access to the Carse (river valley) of Stirling and the approach to the west and north.

If at this time the Castle Rock site was occupied, it would have been by the Wotadini (Votadini), the dominant Celtic tribe of the Lothians, with whom Rome had a relatively stable relationship. The Wotadini capital was on Traprain Law, a cone-shaped hill ("law") some 20 miles east of the modern city; but it appears that in about AD 500, after the Roman withdrawal from Britain, the capital was moved to the site of the present Castle. A poem composed in about AD 600 tells how a band of the Gododdin (an alternative form of Wotadini), from around a place called Din Eidyn, attacked an Anglian force at Catterick in Yorkshire and was annihilated. Din Eidyn—"Eidyn's Hill Fort"—is clearly the Castle Rock site. In the following centuries the Gododdin seem to have been vanquished by Anglian invaders, and by AD 854 Din Eidyn (also spelled Duneideann) had become Edwinesburh (*burh* meaning "stronghold"), giving rise to the supposition that the town was founded by a Northumbrian princeling called Edwin. The date of the poem disproves this, however; it can reasonably be asserted that the Castle Rock site has been continuously occupied for at least 1,400 years, but little is known of the town's earliest centuries.

The Wotadini

THE MEDIEVAL CITY

Mercantile growth. From the mid-11th to the mid-12th century St. Margaret's chapel in the Castle Rock and the Abbey of the Holyrood, endowed by Margaret's son King David I, marked the limits of the Old Town, with the parish church of St. Giles between. In 1130 David I granted Edinburgh the status of a burgh, a privilege

that promoted trade by allowing Edinburgh to act both as a market and as a centre for organized manufacture (particularly of cloth). When David instituted the earliest Scottish coinage, the mint was situated in Edinburgh. He also granted the monks (canons) of Holyrood leave to have their own burgh—with their own jurisdiction; this came to be known as the Canongate. The High Street of Edinburgh ended, and the Canongate began, at the intersection called Netherbow.

Political importance. In 1329 King Robert the Bruce granted Edinburgh a town charter, and its subsequent emergence as the political centre of the nation paralleled its mercantile growth. Perth, a major settlement to the north, had for a time contended for the position of royal capital, but after 1436 Edinburgh's claim was upheld, despite its vulnerability to English invasions. James II (1430–60) was crowned in Holyrood, and most of his parliaments were held in the city (on the site, adjacent to St. Giles, where the Tolbooth was erected in 1466). By the reign of James III (1460–88), Edinburgh was actually described in a royal charter as "the principal burgh of our kingdom" and its established capital.

The medieval burgh—churches and royal and civic buildings apart—was built of wood, and only the houses of the well-to-do had glazed windows and wooden doors. All domestic refuse and the offal of the skinners, butchers, and fishmongers were heaped on either side of the main street, forcing pedestrians to keep to the centre of the street. The graveyards—including that of St. Giles, in the town centre—were used as rubbish dumps. That Edinburgh was uncommonly nasty in this respect was largely the result of its physical setting and the absence of a water supply: until 1681 water had to be fetched from pumped wells south of the Canongate. William Dunbar (*c.* 1460–*c.* 1520), the great Middle Scots poet, wrote in trenchant verse:

> May nane pass through your principal gates
> for stink of haddocks and of skates,
> for cries of carlings [old women] and debates [arguments]. . .
> tailors, souters [shoemakers], and craftis vile
> the fairest of your streets does fyle [defile]. . .

To the moderate prosperity of the late medieval burgh, King James IV (1473–1513) added a touch of European Renaissance culture. He patronized the arts, established in the town the supreme court of justice, and in about 1501 began the construction of a palace at Holyrood, of which only the gatehouse survives. In 1507 in the Cowgate, at the prompting of Bishop Elphinstone, he encouraged the establishment of Scotland's first printing business. In the years of political unrest following the disastrous defeat of the Scots by the English at Flodden in Northumberland (1513), Edinburgh encircled the Old Town (as far as the Netherbow) with a defensive wall, parts of which still stand. It proved sadly ineffective, however, as was shown in 1544 when an English commander, the Earl of Hertford, devastated the entire town, including Canongate and Leith.

Although much of the subsequent rebuilding was still in wood, it was from this period and after that stone became more common, both for public buildings and for the residences of the wealthy. Edinburgh's position made it the seat not only of the court but also of the Privy Council, Parliament, and the law. The great landed families began to keep town houses in the Canongate, and their patronage, frequently more munificent than that of the indigent House of Stuart, stimulated the local economy. Thus, Edinburgh gradually recovered some of the economic and cultural importance lost after Flodden. The Royal College of Surgeons had been founded in 1505. In 1535 the College of Justice was endowed by the church (still at this stage Roman Catholic). And after the young King James VI established himself at Holyrood, he attempted to open Edinburgh up to European culture. In 1582 he granted the town council a charter encouraging the provision of buildings to house the teaching of "humanity, philosophy, theology, medicine, and laws, or of any other liberal sciences whatsoever." This stimulated the foundation of what is today the University of Edinburgh.

Growth in reign of James VI

Disaster and recovery. After 1603, when James VI succeeded to the English throne as James I of England and left for the south, Edinburgh suffered a decline in political and cultural terms, yet the town continued to grow. The first Edinburgh girls' school, the Merchant Maiden Hospital, was opened in 1605, and construction of Heriot's Hospital, a school endowed by the goldsmith and moneylender George Heriot, was begun in 1628. Parliament House, on the site of St. Giles's burying ground, was completed in 1639.

By this time, the area around St. Giles had become the centre of the capital's bustling life. Immediately to the west of the church stood the Tolbooth, combining the roles of council chamber, jail, and place of execution. To the south was the Parliament House and embryonic Parliament Close, with the Court of Exchequer; to the north, the narrow tenement called the Luckenbooths, with its street-level shops. Around the church walls were the *kames*, wooden booths of goldsmiths, jewelers, stationers, and craftsmen. To the east was the Mercat Cross, where business was done from morning to evening.

UNION WITH ENGLAND

Throughout the 17th century the Scottish economy became a weaker relative to England's. In the closing decades of the century, Edinburgh became the head office of an enterprise aimed at establishing a free port at Darién on the Isthmus of Panama, manned by Scots colonists. The scheme failed, however, and by the early 18th century union with England—and thus freedom to trade in the English colonial markets—seemed the last hope of economic growth. In 1707 the Act of Union was signed in a cellar in Parliament Square, and Edinburgh lost all independent political life.

The opportunities for increased trade with a British common market of 7,000,000 people in fact offered little advantage to Edinburgh, for it had no staple manufacture. Yet in the first half of the 18th century it doubled its population and increased its wealth. An observer of 1793 wrote of how "a perpetual influx of the unemployed from the north press into Edinburgh."

Despite the overcrowded conditions in the Old Town, the union brought about a surge of rebuilding and new building within its walls. Much of Parliament Square, badly damaged by fire in 1700, was restored by 1715. New tenement courts were built in the Lawnmarket in the 1720s; the first infirmary (hospital) was opened in 1729, and then a splendid custom-built building was erected in 1748. In the 1730s George Watson's Hospital (a great rival of Heriot's school) was endowed, and in the early 1750s the Royal Exchange (near the City Chambers) was built on the north side of the High Street at the Mercat Cross. It was only after the construction of North Bridge that development took place on a large scale beyond the confines of the Old Town.

Several decades earlier, in the 1720s, the town had reformed and developed its university on the faculty system (the medical faculty was instituted in 1726). This change made possible Edinburgh's contribution to that extraordinary intellectual and cultural flowering known as the Scottish Enlightenment. The Edinburgh Enlightenment itself was very much an Old Town affair. A characteristic of Old Town life was that each multistoried building housed a cross section of Edinburgh society: the very poor at street level or up in the attics; the wealthy on the main floor above street level; and others in between, according to a system whereby the lower the income the less desirable the floor occupied. All shared a common stair and ate and drank in common at the same taverns. Something of this commonness—plainness as well as coarseness—was apparent in Edinburgh's outbreak of intellectual inquiry: a strong, broad, confident ability to grasp the first principles of things and to explain them in the common language, preferably through conversation and debate. So David Hume grasped that there were no uncaused events; Adam Smith recognized the implications of division of labour; Adam Ferguson the danger of "alienation" inherent in labour; William Robertson the degree to which environmental factors shaped economic history; Joseph Black the principle of latent heat; and James Hutton the enormous antiquity of the Earth. As well as these, Edin-

The Edinburgh Enlightenment

burgh was the university of the poet James Thomson; of James Boswell, the biographer of Dr. Johnson; the novelist and poet Oliver Goldsmith; the jurist and writer Lord Kames; the French novelist Benjamin Constant; and Benjamin Rush, an American signatory of the Declaration of Independence.

Toward the end of the 18th century those able to afford the price of a house in the New Town deserted the Old Town. For the first time in five centuries Edinburgh became socially segregated, and, in the political climate of the French Revolution, the city's intellectual elite became authoritarian and antiradical. For 30 years into the 19th century Edinburgh continued to dominate the literary world in Britain, with Sir Walter Scott as its greatest figure, but, by the beginning of Queen Victoria's reign, Edinburgh's intellectual fervour had subsided. The pace of social segregation slowed as well in 1833, when the Town Council, which had sponsored the building of the New Town, went bankrupt.

THE MODERN CITY

From about 1830 until World War I, Edinburgh developed as an industrial centre. A huge growth in the labouring population resulted in severe problems of overcrowding, malnutrition, and epidemics. The city's industries included baking, brewing, distilling, book printing, wire drawing, coach building, and the manufacture of machinery for paper mills along the Water of Leith and the north Esk. The chemical, pharmaceutical, and rubber industries flourished later.

In the 1920s and '30s Edinburgh was at the centre of the Scottish political and literary renaissance led by the nationalist poet Hugh MacDiarmid (pseudonym of Christopher Grieve). Many writers joined in his attempt to revitalize the Lowland Scottish dialect as a literary language. Edinburgh has, since 1930, and particularly since the 1960s, grown more conscious of its Scottish individuality and outlook, seeing this not as parochial and inwardturned but as much more European than the relative isolationism of English culture.

Edinburgh's revival

Four aspects of post-World War II Edinburgh are noteworthy. First is the great expansion of higher education. Since the late 1920s the University of Edinburgh has grown to establish itself as a world leader in areas of advanced research such as medicine and surgery, electronics, and artificial intelligence. Second, the cultural life of the city has expanded, and, although it found major expression in the Edinburgh International Festival, initiated in 1947, firm roots also have been put down in more local enterprises such as the Traverse Theatre, the Demarco art gallery, the restoration by the University of Edinburgh of St. Cecilia's Hall in Cowgate as a small concert hall, the creation of the Scottish Baroque Ensemble, the opening of the Scottish National Gallery of Modern Art, and the conversion of Hope Park Chapel into a home for the Scottish Chamber Orchestra. Literature also has flourished. Small individual publishing firms with an international outlook and a commitment to Scottish writing have reemerged. Third, the city has become acutely conscious of its own heritage in stone and has mounted a strong conservation movement. Local bodies such as the Cockburn and Georgian societies have combined with bodies such as the National Trust for Scotland, the Royal Fine Art Commission, and the Royal Commission on Ancient and Historical Monuments to ensure that the best of the old is preserved and restored and that the worst of the new—including both traffic flow and office blocks—is prevented from intruding into the heart of the city. Finally, Edinburgh has resumed its place as an autonomous political centre. With the establishment of a new Scottish Parliament and government in Edinburgh in 1999, the city has returned to a role as not only the cultural centre but also the political centre of Scotland.

(A.R.T.)

BIBLIOGRAPHY

General: IAN NIMMO, *Portrait of Edinburgh,* 2nd ed. (1975), which describes the city and its inhabitants, their way of life, and culture; GEORGE SCOTT-MONCRIEFF, *Edinburgh,* 3rd ed. (1965), a detailed account of many aspects of the city, illustrated with photographs, prints, and paintings; and ERIC LINKLATER, *Edinburgh* (1960), a personal, essayist treatment of the city.

History: DAVID DAICHES, *Edinburgh* (1978, reissued 1980), which traces the development of the city as a political capital and later a cultural centre; A.J. YOUNGSON, *The Making of Classical Edinburgh: 1750–1840* (1966, reprinted 1975), which details the growth of Edinburgh from a small, crowded town to a substantial and beautiful city; DOUGLAS YOUNG, *Edinburgh in the Age of Sir Walter Scott* (1965), which discusses the city when its vigorous intellectual life gave it the name "Athens of the North"; and E.F. CATFORD, *Edinburgh: The Story of a City* (1975).

Chiefly photographic: DOUGLAS CORRANCE, *Edinburgh,* new ed. (1979), photographs characterizing the principal districts of the city, with captions by W. GORDON SMITH; and A.F. KERSTING, *Portrait of Edinburgh* (1961), mainly architectural features, with text by GEORGE SCOTT-MONCRIEFF.

Special topics: TREVOR ROYLE, *Precipitous City: The Story of Literary Edinburgh* (1980), a history of the city's five centuries as a literary centre; JOHN GIFFORD, COLIN MCWILLIAM, and DAVID WALKER, *Edinburgh* (1984), a comprehensive presentation of the many important buildings in the city; DAVID KEIR (ed.), *The City of Edinburgh,* vol. 15 of *The Third Statistical Account of Scotland* (1966), a close examination of 20th-century life in the city; ROYAL COMMISSION ON THE ANCIENT AND HISTORICAL MONUMENTS AND CONSTRUCTIONS OF SCOTLAND, *An Inventory of the Ancient and Historical Monuments of the City of Edinburgh* (1951), a detailed, illustrated catalog for each district of the city, including a list of monuments the commissioners deemed most worthy of preservation; and ROBERT CHAMBERS, *Traditions of Edinburgh,* 5th ed. (1868, reissued 1980), written to preserve the stories of some of the characters who lived in Edinburgh at the beginning of the 19th century. See also EILEEN DUNLOP and ANTONY KAMM (comps.), *A Book of Old Edinburgh* (1983), an illustrated collection of literary descriptions of the city; and MICHAEL LYNCH, *Edinburgh and the Reformation* (1981), a scholarly treatment of Edinburgh Protestantism, society, and government.

(B.E./A.R.T.)

Edison

Thomas Alva Edison was the quintessential American inventor in the era of Yankee ingenuity. He began his career in 1863, in the adolescence of the telegraph industry, when virtually the only source of electricity was primitive batteries putting out a low-voltage current. Before he died, in 1931, he had played a critical role in introducing the modern age of electricity. From his laboratories and workshops emanated the phonograph, the carbon-button transmitter for the telephone speaker and microphone, the incandescent lamp, a revolutionary generator of unprecedented efficiency, the first commercial electric light and power system, an experimental electric railroad, and key elements of motion-picture apparatus, as well as a host of other inventions. Singly or jointly he held a world-record 1,093 patents. In addition, he created the world's first industrial-research laboratory.

By courtesy of the Edison National Historical Site, West Orange, N.J.



Edison demonstrating his tinfoil phonograph, photograph by Mathew Brady, 1878.

Born in Milan, Ohio, on Feb. 11, 1847, Edison was the seventh and last child—the fourth surviving—of Samuel Edison, Jr., and Nancy Elliot Edison. At an early age he developed hearing problems, which have been variously attributed but were most likely due to a familial tendency to mastoiditis. Whatever the cause, Edison's deafness strongly influenced his behaviour and career, providing the motivation for many of his inventions.

Early years. In 1854 Samuel Edison became the lighthouse keeper and carpenter on the Fort Gratiot military post near Port Huron, Mich., where the family lived in a substantial home. Alva, as the inventor was known until his second marriage, entered school there and attended sporadically for five years. He was imaginative and inquisitive, but because much instruction was by rote and he had difficulty hearing, he was bored and was labeled a misfit. To compensate, he became an avid and omnivorous reader. Edison's lack of formal schooling was not unusual. At the time of the Civil War the average American had attended school a total of 434 days—little more than two years' schooling by today's standards.

In 1859 Edison quit school and began working as a trainboy on the railroad between Detroit and Port Huron. Four years earlier, the Michigan Central had initiated the commercial application of the telegraph by using it to control the movement of its trains, and the Civil War brought a vast expansion of transportation and communication. Edison took advantage of the opportunity to learn telegraphy and in 1863 became an apprentice telegrapher.

Messages received on the initial Morse telegraph were inscribed as a series of dots and dashes on a strip of paper that was decoded and read, so Edison's partial deafness was no handicap. Receivers were increasingly being equipped with a sounding key, however, enabling telegraphers to "read" messages by the clicks. The transformation of telegraphy to an auditory art left Edison more and more disadvantaged during his six-year career as an itinerant telegrapher in the Midwest, the South, Canada, and New England. Amply supplied with ingenuity and insight, he devoted much of his energy toward improving the inchoate equipment and inventing devices to facilitate some of the tasks that his physical limitations made difficult. By January 1869 he had made enough progress with a duplex telegraph (a device capable of transmitting two messages simultaneously on one wire) and a printer, which converted electrical signals to letters, that he abandoned telegraphy for full-time invention and entrepreneurship.

Early inventions

Edison moved to New York City, where he initially went into partnership with Frank L. Pope, a noted electrical expert, to produce the Edison Universal Stock Printer and other printing telegraphs. Between 1870 and 1875 he worked out of Newark, N.J., and was involved in a variety of partnerships and complex transactions in the fiercely competitive and convoluted telegraph industry, which was dominated by the Western Union Telegraph Company. As an independent entrepreneur he was available to the highest bidder and played both sides against the middle. During this period he worked on improving an automatic telegraph system for Western Union's rivals. The automatic telegraph, which recorded messages by means of a chemical reaction engendered by the electrical transmissions, proved of limited commercial success, but the work advanced Edison's knowledge of chemistry and laid the basis for his development of the electric pen and mimeograph, both important devices in the early office machine industry, and indirectly led to the discovery of the phonograph. Under the aegis of Western Union he devised the quadruplex, capable of transmitting four messages simultaneously over one wire, but railroad baron and Wall Street financier Jay Gould, Western Union's bitter rival, snatched the quadruplex from the telegraph company's grasp in December 1874 by paying Edison more than $100,000 in cash, bonds, and stock, one of the larger payments for any invention up to that time. Years of litigation followed.

Menlo Park. Although Edison was a sharp bargainer, he was a poor financial manager, often spending and giving away money more rapidly than he earned it. In 1871 he married 16-year-old Mary Stilwell, who was as improvident in household matters as he was in business, and before the end of 1875 they were in financial difficulties. To reduce his costs and the temptation to spend money, Edison brought his now-widowed father from Port Huron to build a 2½-story laboratory and machine shop in the rural environs of Menlo Park, N.J.—12 miles south of Newark—where he moved in March 1876. Accompanying him were two key associates, Charles Batchelor and John Kruesi. Batchelor, born in Manchester in 1845, was a master mechanic and draftsman who complemented Edison perfectly and served as his "ears" on such projects as the phonograph and telephone. He was also responsible for fashioning the drawings that Kruesi, a Swiss-born machinist, translated into models.

Financial problems

Edison experienced his finest hours at Menlo Park. While experimenting on an underwater cable for the automatic telegraph, he found that the electrical resistance and conductivity of carbon (then called plumbago) varied according to the pressure it was under. This was a major theoretical discovery, which enabled Edison to devise a "pressure relay" using carbon rather than the usual mag-

nets to vary and balance electric currents. In February 1877 Edison began experiments designed to produce a pressure relay that would amplify and improve the audibility of the telephone, a device that Edison and others had studied but which Alexander Graham Bell was the first to patent, in 1876. By the end of 1877 Edison had developed the carbon-button transmitter that is still used in telephone speakers and microphones.

Edison invented many items, including the carbon transmitter, in response to specific demands for new products or improvements. But he also had the gift of serendipity: when some unexpected phenomenon was observed, he did not hesitate to halt work in progress and turn off course in a new direction. This was how, in 1877, he achieved his most original discovery, the phonograph. Because the telephone was considered a variation of acoustic telegraphy, Edison during the summer of 1877 was attempting to devise for it, as he had for the automatic telegraph, a machine that would transcribe signals as they were received, in this instance in the form of the human voice, so that they could then be delivered as telegraph messages. (The telephone was not yet conceived as a general, person-to-person means of communication.) Some earlier researchers, notably the French inventor Léon Scott, had theorized that each sound, if it could be graphically recorded, would produce a distinct shape resembling shorthand, or phonography ("sound writing"), as it was then known. Edison hoped to reify this concept by employing a stylus-tipped carbon transmitter to make impressions on a strip of paraffined paper. To his astonishment, the scarcely visible indentations generated a vague reproduction of sound when the paper was pulled back beneath the stylus.

The phonograph

Edison unveiled the tinfoil phonograph, which replaced the strip of paper with a cylinder wrapped in tinfoil, in December 1877. It was greeted with incredulity. Indeed, a leading French scientist declared it to be the trick device of a clever ventriloquist. The public's amazement was quickly followed by universal acclaim. Edison was projected into worldwide prominence and was dubbed the Wizard of Menlo Park, although a decade passed before the phonograph was transformed from a laboratory curiosity into a commercial product.

Another offshoot of the carbon experiments reached fruition sooner. Samuel Langley, Henry Draper, and other American scientists needed a highly sensitive instrument that could be used to measure minute temperature changes in heat emitted from the Sun's corona during a solar eclipse along the Rocky Mountains on July 29, 1878. To satisfy those needs Edison devised a "microtasimeter" employing a carbon button. This was a time when great advances were being made in electric arc lighting, and during the expedition, which Edison accompanied, the men discussed the practicality of "subdividing" the intense arc lights so that electricity could be used for lighting in the same fashion as with small, individual gas "burners." The basic problem seemed to be to keep the burner, or bulb, from being consumed by preventing it from overheating. Edison thought he would be able to solve this by fashioning a microtasimeter-like device to control the current. He boldly announced that he would invent a safe, mild, and inexpensive electric light that would replace the gaslight.

The electric light

The incandescent electric light had been the despair of inventors for 50 years, but Edison's past achievements commanded respect for his boastful prophecy. Thus, a syndicate of leading financiers, including J.P. Morgan and the Vanderbilts, established the Edison Electric Light Company and advanced him $30,000 for research and development. Edison proposed to connect his lights in a parallel circuit by subdividing the current, so that, unlike arc lights, which were connected in a series circuit, the failure of one light bulb would not cause a whole circuit to fail. Some eminent scientists predicted that such a circuit could never be feasible, but their findings were based on systems of lamps with low resistance—the only successful type of electric light at the time. Edison, however, determined that a bulb with high resistance would serve his purpose, and he began searching for a suitable one.

He had the assistance of 26-year-old Francis Upton, a graduate of Princeton University with an M.A. in science. Upton, who joined the laboratory force in December 1878, provided the mathematical and theoretical expertise that Edison himself lacked. (Edison later revealed, "At the time I experimented on the incandescent lamp I did not understand Ohm's law." On another occasion he said, "I do not depend on figures at all. I try an experiment and reason out the result, somehow, by methods which I could not explain.")

By the summer of 1879 Edison and Upton had made enough progress on a generator—which, by reverse action, could be employed as a motor—that Edison, beset by failed incandescent lamp experiments, considered offering a system of electric distribution for power, not light. By October Edison and his staff had achieved encouraging results with a complex, regulator-controlled vacuum bulb with a platinum filament, but the cost of the platinum would have made the incandescent light impractical. While experimenting with an insulator for the platinum wire, they discovered that, in the greatly improved vacuum they were now obtaining through advances made in the vacuum pump, carbon could be maintained for some time without elaborate regulatory apparatus. Advancing on the work of Joseph Wilson Swan, an English physicist, Edison found that a carbon filament provided a good light with the concomitant high resistance required for subdivision. Steady progress ensued from the first breakthrough in mid-October until the initial demonstration for the backers of the Edison Electric Light Company on December 3.

It was, nevertheless, not until the summer of 1880 that Edison determined that carbonized bamboo fibre made a satisfactory material for the filament, although the world's first operative lighting system had been installed on the steamship *Columbia* in April. The first commercial land-based "isolated" (single-building) incandescent system was placed in the New York printing firm of Hinds and Ketcham in January 1881. In the fall a temporary, demonstration central power system was installed at the Holborn Viaduct in London, in conjunction with an exhibition at the Crystal Palace. Edison himself supervised the laying of the mains and installation of the world's first permanent, commercial central power system in lower Manhattan, which became operative in September 1882. Although the early systems were plagued by problems and many years passed before incandescent lighting powered by electricity from central stations made significant inroads into gas lighting, isolated lighting plants for such enterprises as hotels, theatres, and stores flourished—as did Edison's reputation as the world's greatest inventor.

The world's first electric lighting system

One of the accidental discoveries made in the Menlo Park laboratory during the development of the incandescent light anticipated the British physicist J.J. Thomson's discovery of the electron 15 years later. In 1881–82 William J. Hammer, a young engineer in charge of testing the light globes, noted a blue glow around the positive pole in a vacuum bulb and a blackening of the wire and the bulb at the negative pole. This phenomenon was first called "Hammer's phantom shadow," but when Edison patented the bulb in 1883 it became known as the "Edison effect." Scientists later determined that this effect was explained by the thermionic emission of electrons from the hot to the cold electrode, and it became the basis of the electron tube and laid the foundation for the electronics industry.

Edison had moved his operations from Menlo Park to New York City when work commenced on the Manhattan power system. Increasingly, the Menlo Park property was used only as a summer home. In August 1884 Edison's wife, Mary, suffering from deteriorating health and subject to periods of mental derangement, died there of "congestion of the brain," apparently a tumour or hemorrhage. Her death and the move from Menlo Park roughly mark the halfway point of Edison's life.

The Edison Laboratory. A widower with three young children, Edison, on Feb. 24, 1886, married 20-year-old Mina Miller, the daughter of a prosperous Ohio manufacturer. He purchased a hilltop estate in West Orange, N.J., for his new bride and constructed nearby a grand, new laboratory, which he intended to be the world's first true research facility. There, he produced the commercial phonograph, founded the motion-picture industry, and

developed the alkaline storage battery. Nevertheless, Edison was past the peak of his productive period. A poor manager and organizer, he worked best in intimate, relatively unstructured surroundings with a handful of close associates and assistants; the West Orange laboratory was too sprawling and diversified for his talents. Furthermore, as a significant portion of the inventor's time was taken up by his new role of industrialist, which came with the commercialization of incandescent lighting and the phonograph, electrical developments were passing into the domain of university-trained mathematicians and scientists. Above all, for more than a decade Edison's energy was focused on a magnetic ore-mining venture that proved the unquestioned disaster of his career.

The first major endeavour at the new laboratory was the commercialization of the phonograph, a venture launched in 1887 after Alexander Graham Bell, his cousin Chichester, and Charles Tainter had developed the graphophone—an improved version of Edison's original device—which used waxed cardboard instead of tinfoil. Two years later, Edison announced that he had "perfected" the phonograph, although this was far from true. In fact, it was not until the late 1890s, after Edison had established production and recording facilities adjacent to the laboratory, that all the mechanical problems were overcome and the phonograph became a profitable proposition.

In the meantime, Edison conceived the idea of popularizing the phonograph by linking to it in synchronization a zoetrope, a device that gave the illusion of motion to photographs shot in sequence. He assigned the project to William K.L. Dickson, an employee interested in photography, in 1888. After studying the work of various European photographers who also were trying to record motion, Edison and Dickson succeeded in constructing a working camera and a viewing instrument, which were called, respectively, the Kinetograph and the Kinetoscope. Synchronizing sound and motion proved of such insuperable difficulty, however, that the concept of linking the two was abandoned, and the silent movie was born. Edison constructed at the laboratory the world's first motion-picture stage, nicknamed the "Black Maria," in 1893, and the following year Kinetoscopes, which had peepholes that allowed one person at a time to view the moving pictures, were introduced with great success. Rival inventors soon developed screen-projection systems that hurt the Kinetoscope's business, however, so Edison acquired a projector developed by Thomas Armat and introduced it as "Edison's latest marvel, the Vitascope."

Motion pictures

Another derivative of the phonograph was the alkaline storage battery, which Edison began developing as a power source for the phonograph at a time when most homes still lacked electricity. Although it was 20 years before all the difficulties with the battery were solved, by 1909 Edison was a principal supplier of batteries for submarines and electric vehicles and had even formed a company for the manufacture of electric automobiles. In 1912 Henry Ford, one of Edison's greatest admirers, asked him to design a battery for the self-starter, to be introduced on the Model T. Ford's request led to a continuing relationship between these two Americans, and in October 1929 he staged a 50th-anniversary celebration of the incandescent light that turned into a universal apotheosis for Edison.

Most of Edison's successes involved electricity or communication, but throughout the late 1880s and early 1890s the Edison Laboratory's top priority was the magnetic ore-separator. Edison had first worked on the separator when he was searching for platinum for use in the experimental incandescent lamp. The device was supposed to cull platinum from iron-bearing sand. During the 1880s iron ore prices rose to unprecedented heights, so that it appeared that, if the separator could extract the iron from unusable low-grade ores, then abandoned mines might profitably be placed back in production. Edison purchased or acquired rights to 145 old mines in the east and established a large pilot plant at the Ogden mine, near Ogdensburg, N.J. He was never able to surmount the engineering problems or work the bugs out of the system, however, and when ore prices plummeted in the mid-1890s he gave up on the idea. By then he had liquidated all but a small part of his holdings in the General Electric Company, sometimes at very low prices, and had become more and more separated from the electric lighting field.

The magnetic ore-separator

Failure could not discourage Edison's passion for invention, however. Although none of his later projects were as successful as his earlier ones, he continued to work even in his 80s. He died in West Orange on Oct. 18, 1931.

Assessment. The thrust of Edison's work may be seen in the clustering of his patents: 389 for electric light and power, 195 for the phonograph, 150 for the telegraph, 141 for storage batteries, and 34 for the telephone. His life and achievements epitomize the ideal of applied research. He always invented for necessity, with the object of devising something new that he could manufacture. The basic principles he discovered were derived from practical experiments, invariably by chance, thus reversing the orthodox concept of pure research leading to applied research.

Edison's role as a machine shop operator and small manufacturer was crucial to his success as an inventor. Unlike other scientists and inventors of the time, who had limited means and lacked a support organization, Edison ran an inventive establishment. He was the antithesis of the lone inventive genius, although his deafness enforced on him an isolation conducive to conception. His lack of managerial ability was, in an odd way, also a stimulant. As his own boss, he plunged ahead on projects more prudent men would have shunned, then tended to dissipate the fruits of his inventiveness, so that he was both free and forced to develop new ideas. Few men have matched him in the positiveness of his thinking. Edison never questioned whether something might be done, only how.

Edison's career, the fulfillment of the American dream of rags-to-riches through hard work and intelligence, made him a folk hero to his countrymen. In temperament he was an uninhibited egotist, at once a tyrant to his employees and their most entertaining companion, so that there was never a dull moment with him. He was charismatic and courted publicity, but he had difficulty socializing and neglected his family. His shafts at the expense of the "long-haired" fraternity of theorists sometimes led formally trained scientists to deprecate him as anti-intellectual; yet he employed as his aides, at various times, a number of eminent mathematical physicists, such as Nikola Tesla and A.E. Kennelly. The contradictory nature of his forceful personality, as well as such eccentricities as his ability to catnap anywhere, contributed to his legendary status. By the time he was in his middle 30s Edison was said to be the best-known American in the world. When he died he was venerated and mourned as the man who, more than any other, had laid the basis for the technological and social revolution of the modern electric world.

BIBLIOGRAPHY. ALFRED O. TATE, *Edison's Open Door: The Life Story of Thomas A. Edison, a Great Individualist* (1938), which tells the story of the early years of the West Orange laboratory, was written by Edison's secretary of the period. FRANCIS JEHL, *Menlo Park Reminiscences,* 3rd ed. (1937–41), is a firsthand account of the 1878–80 period at Menlo Park, by an assistant who came to dislike Edison but was later the first curator at Henry Ford's Edison Institute. THOMAS A. EDISON, *The Diary and Sundry Observations of Thomas A. Edison,* ed. by DAGOBERT D. RUNES (1948, reprinted 1976), provides insight into Edison's feelings and thoughts, especially in the period following the death of his first wife. MATTHEW JOSEPHSON, *Edison: A Biography* (1959), is based on the correspondence and laboratory notebooks in the Edison Laboratory archives, though at the time of its publication the access to the records was severely restricted, which makes the book outdated. ROBERT CONOT, *A Streak of Luck* (1979, reprinted 1986 as *Thomas A. Edison*), is the first comprehensive biography based entirely on the original sources from the West Orange and other depository archives. WYN WACHHORST, *Thomas Alva Edison: An American Myth* (1981), is a revisionist study of Edison's place in the cultural history of the United States, with an extensive bibliography. See also ROBERT FRIEDEL and PAUL ISRAEL, *Edison's Electric Light: A Biography of an Invention* (1986), a well-researched, illustrated account. Archival papers of Edison and his associates are published in *Thomas A. Edison Papers: A Selective Microfilm Edition* (1985–); part 1, for the period 1850–78, and part 2, for 1879–86, have been filmed from the West Orange archives. Subsequent parts will include documents from other repositories.

(M.Jo./R.E.Co.)